MSIN0096 Homework 4
Due Dec 6th, 3pm

1. (21pt) We have 4 independent variables X_1, X_2, X_3 and X_4. The table as below provides R^2, adjusted R^2 and BIC of 15 different OLS models. Column"Independent variables" includes the set of independent variables contained in each OLS regression.

Independent variables	R^2	Adj R^2	BIC
X_1	0.0011	0.0010	823.049
X_2	0.1616	0.1574	788.009
X_3	0.5274	0.5250	673.3828
X_4	0.0047	0.0040	822.3285
X_1,X_2	0.1618	0.1533	793.2606
X_1,X_3	0.5342	0.5294	675.7815
X_1,X_4	0.0059	0.0050	827.3785
X_2,X_3	0.6766	0.6733	602.8175
X_2,X_4	0.1675	0.1590	791.9065
X_3,X_4	0.5426	0.5379	672.1306
X_1,X_2,X_3	0.6806	0.6757	605.5908
X_1,X_2,X_4	0.1677	0.1550	797.1405
X_1,X_3,X_4	0.5502	0.5433	674.084
X_2,X_3,X_4	0.6936	0.6889	597.2843
X_1,X_2,X_3,X_4	0.6983	0.6921	599.4996

   (a) (7pt) Use best subset selection and BIC criterion to choose the best model. Explain each step of your procedure.

   (b) (7pt) Use forward stepwise selection and BIC criterion to choose the best model. Explain each step of your procedure.

   (c) (7pt) Use backward stepwise selection and the BIC criterion to choose the best model. Explain each step of your procedure.

2. (29pt) In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

   (a) (3pt) Use the *rnorm*() function to generate a variable X of length $n = 100$, as well as a noise vector e of length n = 100.

   (b) (2pt) Generate a response vector Y of length $n = 100$ according to the model as below.

$$Y = 1 + 2X + 3X^2 + 4X^3 + e$$

   Note that this is the underlying true model.

   (c) (6pt) Use the *regsubsets*() function to perform best subset selection in order to choose the best model containing the variables $X, X^2, \cdots, X^{10}$. What is the best model obtained according to BIC? Report the coefficients of the best model obtained.

   (d) (6pt) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

(e) (6pt) Now fit a ridge regression to the simulated data, again using $X, X^2, \cdots, X^{10}$ as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates.

(f) (6pt) Repeat (e), but now fit a lasso regression. Compare the coefficients with the results in (e).

3. (30pt) This question involves the dataset Q3.csv which contains following variables.

$Purchase$: CH and MM indicating whether the customer purchased Citrus Hill(CH) or Minute Maid(MM) Orange Juice
$WeekofPurchase$: Week of purchase
$StoreID$: Store ID
$PriceCH$: Price charged for CH
$PriceMM$: Price charged for MM
$DiscCH$: Discount offered for CH
$DiscMM$: Discount offered for MM
$SpecialCH$: Indicator of special on CH
$SpecialMM$: Indicator of special on MM
$LoyalCH$: Customer brand loyalty for CH
$SalePriceMM$: Sale price for MM
$SalePriceCH$: Sale price for CH
$PriceDiff$: Sale price of MM less sale price of CH
$Store7$: A factor with levels No and Yes indicating whether the sale is at Store 7
$PctDiscMM$: Percentage discount for MM
$PctDiscCH$: Percentage discount for CH
$ListPriceDiff$: List price of MM less list price of CH
$STORE$ Which of 5 possible stores the sale occured at

(a) (3pt) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

(b) (6pt) Fit a tree to the training data, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics about the tree. What is the training error rate? How many terminal nodes does the tree have?

(c) (5pt) Predict the response on the test data. What is the test error rate?

(d) (6pt) Apply the cv.tree() function to the training set in order to determine the optimal tree size. Which tree size corresponds to the lowest cross-validated classification error rate? Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If there is a tie on cross-validation errors, prune the tree to the smallest size.

(e) (5pt) Fit random forests using the training data and list the three most important variables to explain the purchase of orange juice.

(f) (5pt) Use the test data to compute the test errors of the unpruned trees, pruned tree and random forests. Which is the lowest?

4. (20pt) The table below provides the similarity matrix of 6 observations and each cell in this table is the Euclidean distance between 2 observations. Please use the bottom up (agglomera-

tive) approach and the complete linkage to build the hierarchical clustering manually. Clearly describe and each step and draw the dendrogram (or the clustering tree).

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0					
X_2	5.2	0				
X_3	1.1	4.1	0			
X_4	4.5	1.5	3.7	0		
X_5	5.5	3.8	3.9	3.6	0	
X_6	6.4	3.5	4.8	5.5	2.1	0