ISL Fall 2015 Assignment V. 100 pts.

PIN NUMBER(s):

**LOKESHWAR REDDY INJA (16207239/ LIRK9)**

**SANTHOSH KUMAR GATTU (16211118/ SG6N6)**

**SANTOSH KUMAR (16207971/ SK7Z9)**

You may submit this assignment in groups of upto three each. Write your **PIN numbers** on this sheet and include it as the cover page for your submission.

The objective of this assignment is to practice using R, and gain a fundamental understanding of linear regression. Your submission should include both your code as well your answers to the questions.

Electronic submission on Blackboard is due **latest by 11 pm on Wed, Nov 4th**. You may upload upto **three** submissions **before** the deadline – only the last submission will be graded. Submissions received after the deadline will be graded only for effort for a maximum of 70% of the total grade (Refer to class syllabus for detailed grading policy).

**State any assumptions you make, justify your answers, show intermediate steps and explain your results for maximum credit**. All answers should be in your own words with any sources you refer to cited at the appropriate places. Any knowledge you acquire from the Internet should be written in your own words and be appropriately referenced. Copying and pasting from the Internet, each other or any other source will not count as your effort (Refer to class syllabus for detailed policy on plagiarism).

**Remember that answers need to be word-processed (NOT handwritten) and should use R. Submit all your R code as a single merged file for all the assignments.**

Answer the following questions from Chapter 6.

Q8abcd

Q9

**8. In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.**

**(a)  Use the rnorm() function to generate a predictor X of length n = 100, as well as a noise vector ε of length n = 100.**

**Sol:** set.seed(1)

X=rnorm(100,mean=0,sd=1)

e=rnorm(100,mean=0,sd=0.5)

**(b)  Generate a response vector Y of length n = 100 according to the model  Y = β0 +β1X +β2X2 +β3X3 +ε, where β0, β1, β2, and β3 are constants of your choice.**

**Sol:** B0=150.3

B1=50.5

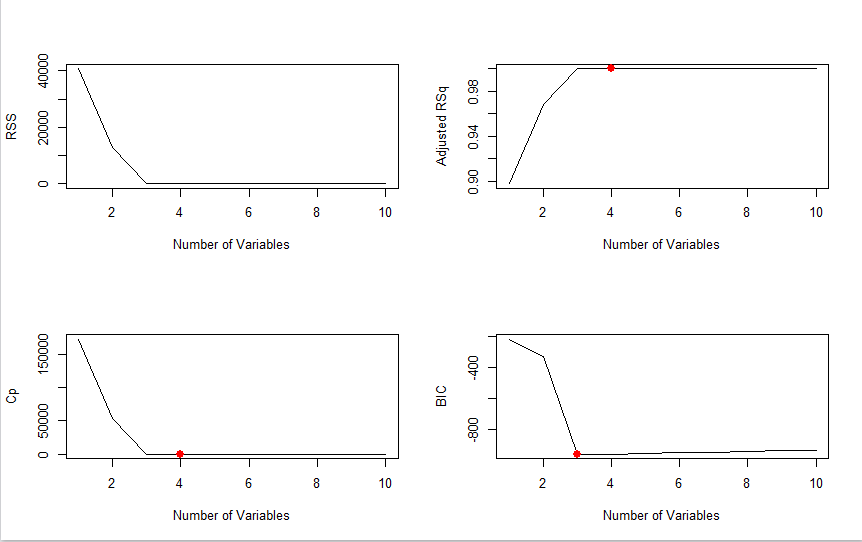
B2=-10.1

B3=-34.2

Y=B0+ B1\*X + B2\*(X^2) + B3\*(X^3) + e

**(c)  Use the regsubsets() function to perform best subset selection in order to choose the best model containing the predictors X,X2,...,X10. What is the best model obtained according to Cp, BIC, and adjusted R2? Show some plots to provide evidence for your answer, and report the coefficients of the best model ob- tained. Note you will need to use the data.frame() function to create a single data set containing both X and Y .**

**Sol.** R code is attached separately.



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

**Sol.** R code is attached separately.

Same number of parameters were used by above methods but some false parameters changed with Cp and Adjusted R^2. x^5 is used as extra by Cp and Adjusted R^2.

**9. In this exercise, we will predict the number of applications received using the other variables in the College data set.**

**(a). Split the data set into a training set and a test set.**

Sol: R code is attached separately

**(b). Fit a linear model using least squares on the training set, and  report the test error obtained.**

Sol: R code is attached separately.

Test Error obtained is 0.9051744.

.

**(c).Fit a ridge regression model on the training set, with λ chosen  by cross-validation. Report the test error obtained.**

**Sol:**

Test Error obtained is 0.8386754.

.

R code is attached separately.

**(d)**

**Sol:**

Test Error obtained is 0.8999059.

Number of non-zero coefficient estimates are 3.

R code is attached separately.

**(e)**

**Sol:**

Test Error obtained is 0.9051744.

Number of components considered is 17.

Cross validation is done using 10 random segments.

**(f)**

**Sol:**

Test Error obtained is 0.9034921

Number of components considered is 17.

Cross validation is done using 10 random segments.

**(g)**

**Sol:**

Best performed models are Ordinary least squares, PLS regression, lasso, and PCR regression as these methods use the same number of variables.

PLS regression model used 9 out of 17 components and 83% of the variance thus performing well to some extent.

Lasso only made 3 variables to 0 and it did not perform well

Lease performed model is Ridge regression.