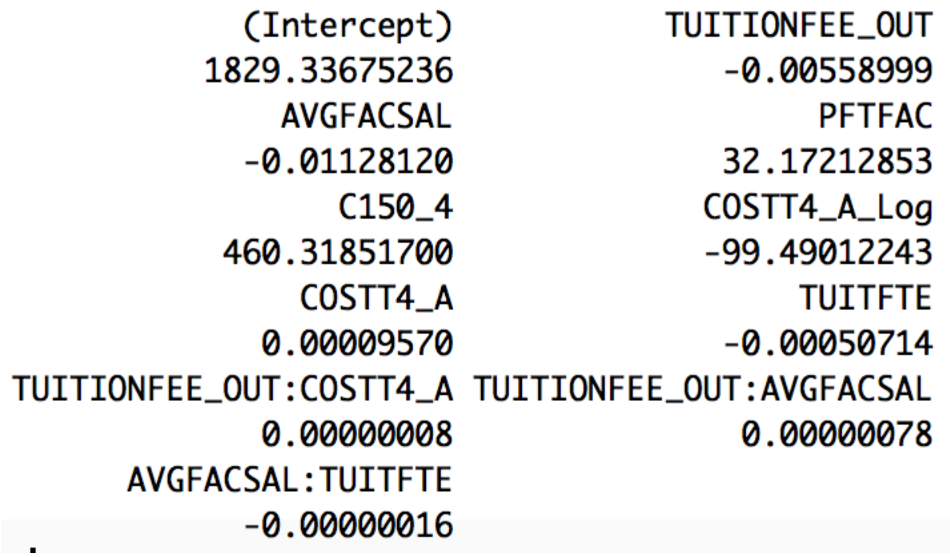
Name: Parth Pareek
UNI: PP2547
Date: 02/15/2016
Assignment: HW2

1. Number of rows after removing NAs – 1136
2. Number of covariates (predictive variables / number of columns) – 19

SAT_AVG	UGDS	COSTT4_A	TUITIONFEE_OUT	TUITFTE
1066.26	6292.74	31929.56	24963.24	12468.33
C150_4	PFTFTUG1_EF	COSTT4_A_Log	TUITIONFEE_OUT_Log	TUITFTE_Log
0.55	0.63	10.29	10.05	9.29
AVGFACSAL	PFTFAC	AVGFACSAL_Log	COSTT4_A* TUITIONFEE_OUT	COSTT4_A* TUITFTE
7685.10	0.71	8.91	899966138.71	469864196.14
COSTT4_A* AVGFACSAL	TUITIONFEE_OUT* TUITFTE	TUITIONFEE_OUT* AVGFACSAL	TUITFTE* AVGFACSAL	
253594680.83	360243222.62	201723570.86	101550437.18	

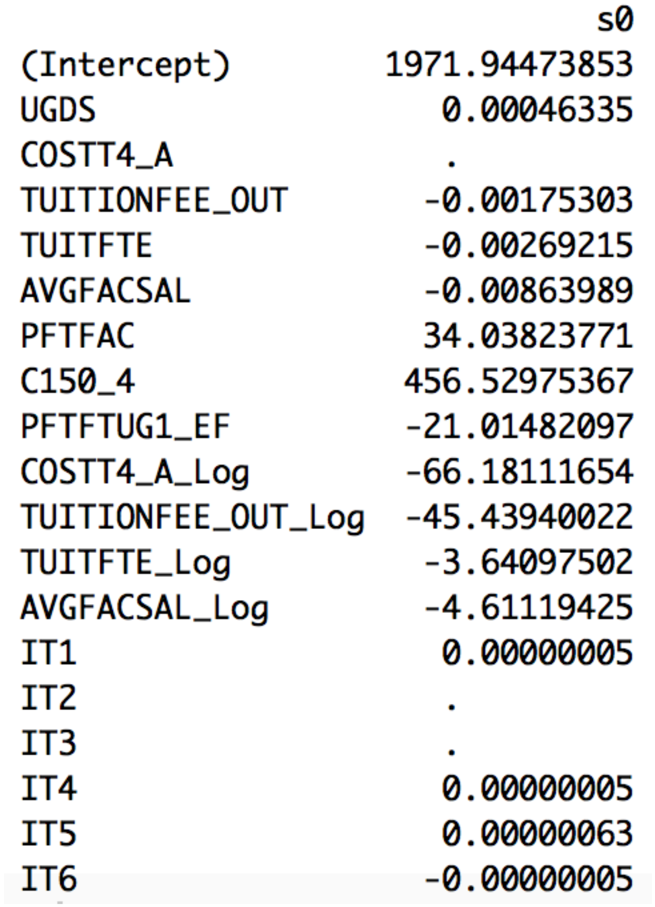
3. SAT_AVG for training data set - 1066.904
   SAT_AVG for test data set - 1064.331

4. Subset size 8 is best. Model parameters - TUITIONFEE_OUT, AVGFACSAL, PFTFAC, C150_4, COSTT4_A_Log, COSTT4_A*TUITIONFEE_OUT, TUITIONFEE_OUT*AVGFACSAL, TUITFTE*AVGFACSAL



```
          (Intercept)          TUITIONFEE_OUT
        1829.33675236             -0.00558999
            AVGFACSAL                  PFTFAC
          -0.01128120             32.17212853
               C150_4            COSTT4_A_Log
         460.31851700            -99.49012243
             COSTT4_A                 TUITFTE
           0.00009570             -0.00050714
TUITIONFEE_OUT:COSTT4_A TUITIONFEE_OUT:AVGFACSAL
           0.00000008              0.00000078
      AVGFACSAL:TUITFTE
          -0.00000016
```

5. Lambda for best model: 0.1



```
                                 s0
(Intercept)       1971.94473853
UGDS                 0.00046335
COSTT4_A             .
TUITIONFEE_OUT      -0.00175303
TUITFTE             -0.00269215
AVGFACSAL           -0.00863989
PFTFAC              34.03823771
C150_4             456.52975367
PFTFTUG1_EF        -21.01482097
COSTT4_A_Log       -66.18111654
TUITIONFEE_OUT_Log -45.43940022
TUITFTE_Log         -3.64097502
AVGFACSAL_Log       -4.61119425
IT1                  0.00000005
IT2                  .
IT3                  .
IT4                  0.00000005
IT5                  0.00000063
IT6                 -0.00000005
```

IT1…IT6 refers to the interaction terms
IT1: COSTT4_A* TUITIONFEE_OUT
IT2: COSTT4_A* TUITFTE
IT3: COSTT4_A* AVGFACSAL
IT4: TUITIONFEE_OUT* TUITFTE
IT5: TUITIONFEE_OUT* AVGFACSAL
IT6: TUITFTE* AVGFACSAL

6. We choose the model from Best Subset Selection since it has lower MSE.
   MSE for model from Best Subset selection is 5301.185
   MSE for model from Lasso regression is 5344.171

7. MSE from best subset selection was lower and was chosen as the best model. However, the maximum limit for subset size was specified to be 8. I will try to increase the maximum limit and see if there is a change. Also, interaction term TUITIONFEE_OUT* TUITFTE is part of Lasso subset and not subset selection. You need to delve deeper into this issue as well. Subset selection includes some interaction terms. Of them, the original terms are not included in the regression (e.g. COSTT4_A). This is not the best practice and if the term were included, it may result in a higher MSE. This also needs to be looked into.