COLLEGE GRADUATION RATES: ANALYSIS OF DATA

LINEAR REGRESSION PROJECT: REPORT I

By

Parag Shah

INTRODUCTION

Problem Description and objective:

Question that I am trying to answer is "whether there exist linear relationship among various factors and student graduation rate and what are the most important factors affecting it"

Data Description:

The data set is drawn from two sources, U.S. News & World Report's "Guide to Americas Best colleges" and AAUP 1994 Salary survey. There are 31 quantitative variables and 4 qualitative variables which are not going to be included for doing analysis. The response variable, Graduation Rate, is quantitative.

The U.S. News data contains information on tuition, room & board costs, SAT or ACT scores, application/acceptance rates, graduation rate, student/faculty ratio, spending per student, and a number of other variables for 1300+ schools. The AAUP data includes average salary, overall compensation, and number of faculty broken down by full, associate, and assistant professor ranks.

Report Objective:

Objective of first report is identifying correlated related variables, provide descriptive statistics, detecting colinearity among them and provide analysis of each variable and response variable.

DATA ANALYSIS:

General statistics:

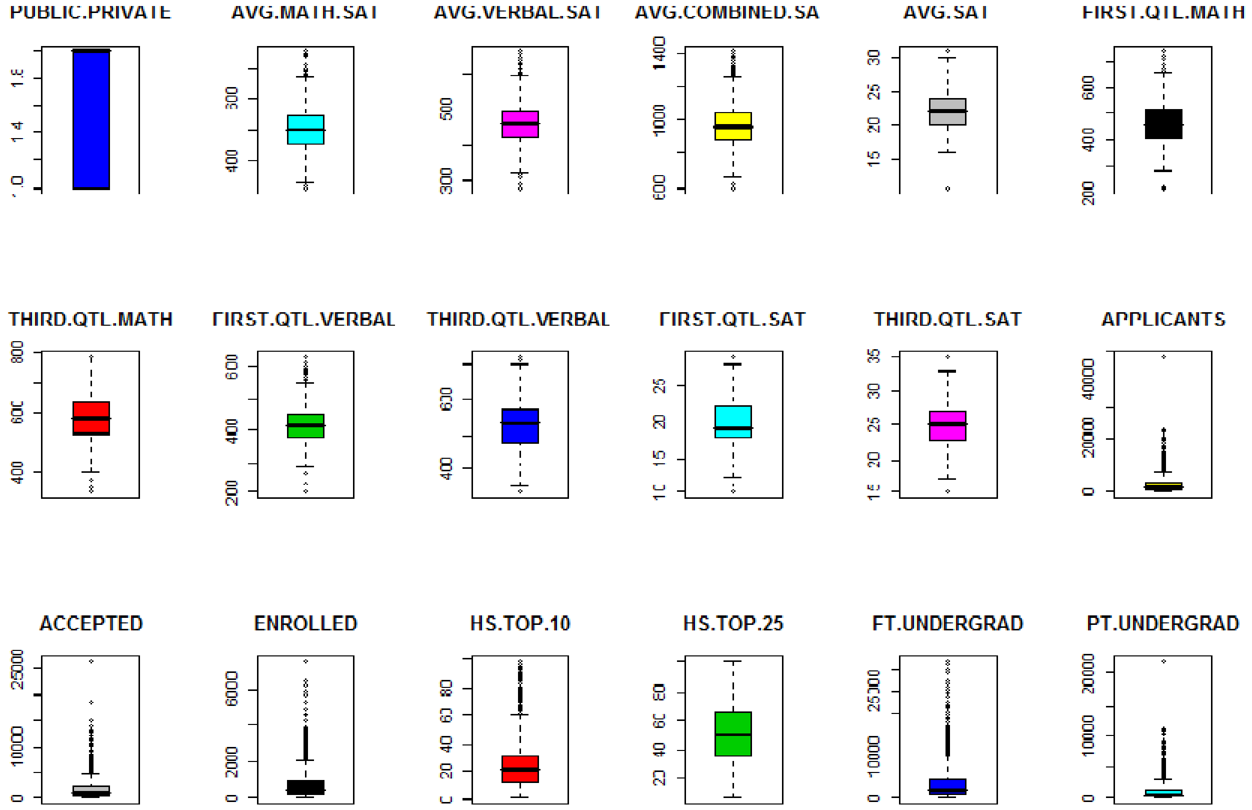
Before I start to build a model for the graduation rates, here are some general statistics that are associated with colleges' data. From table below, some of the data contains erroneous value e.g. graduation rate of 118 and max combined SAT of 1410. These are some of the possible outliers which I have to deal with.

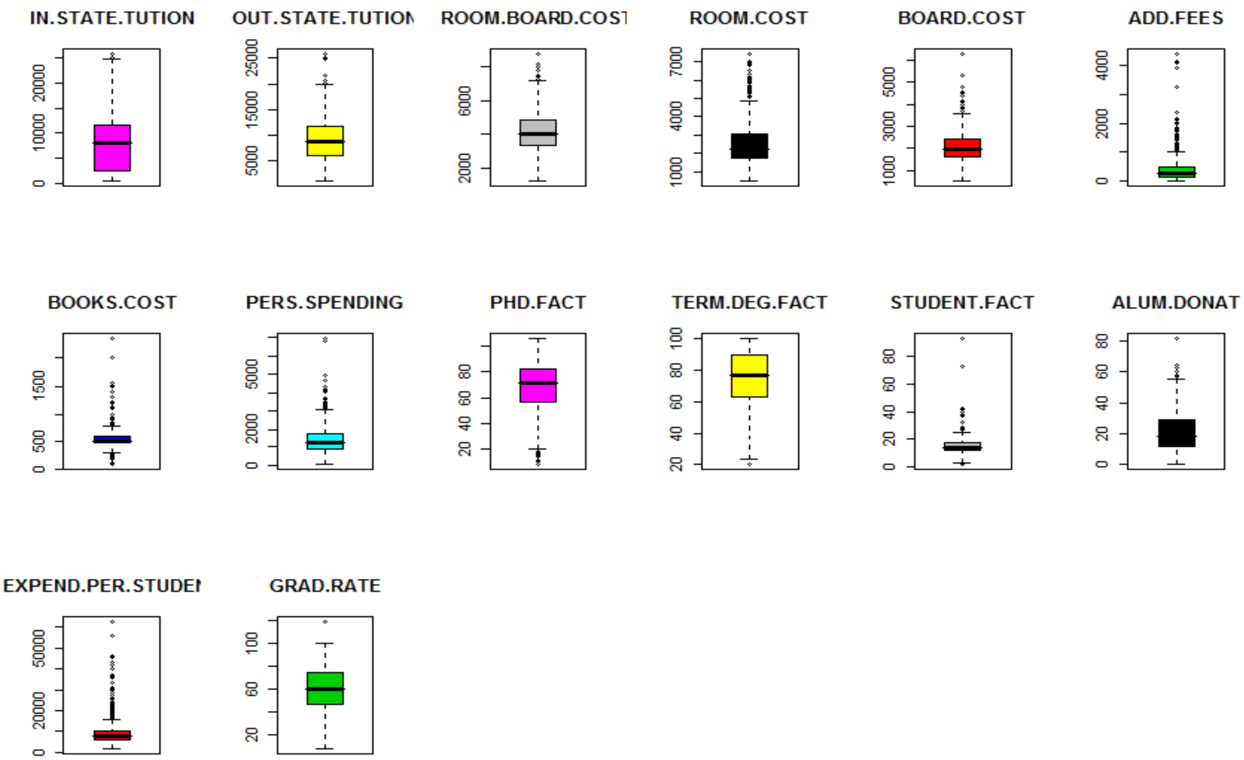
Five number summary of each input variable

	Minimum	Lower Hinge	Median	Upper Hinge	Maximum	Mean	median	Std.dev
PUBLIC.PRIVATE	1	1	2	2	2	1.639017	2	0.48047
AVG.MATH.SAT	320	460	500	544	750	506.8378	500	67.82244
AVG.VERBAL.SAT	280	422	457	492	665	461.2239	457	58.29841
AVG.COMBINED.SAT	600	884.5	957	1038	1410	967.9782	957	123.5775
AVG.SAT	11	20	22	24	31	22.12045	22	2.579899
FIRST.QTL.MATH	220	410	453	510	740	462.2358	453	76.32264
THIRD.QTL.MATH	330	530	580	630	785	583.149	580	71.21797
FIRST.QTL.VERBAL	200	380	410	450	630	418.487	410	64.49204
THIRD.QTL.VERBAL	330	480	530	570	720	530.4521	530	64.53681
FIRST.QTL.SAT	10	18	19	22	29	19.819	19	2.796275
THIRD.QTL.SAT	15	23	25	27	35	25.11312	25	2.781172
APPLICANTS	35	695.5	1470	3314.5	48094	2752.098	1470	3541.975
ACCEPTED	35	554.5	1095	2303	26330	1870.683	1095	2250.866
ENROLLED	18	236	447	984	7425	778.8805	447	884.5783
HS.TOP.10	1	13	21	32	98	25.67198	21	18.31262
HS.TOP.25	6	36.5	50	66	100	52.35	50	20.88132
FT.UNDERGRAD	59	966	1812	4539.5	31643	3692.665	1812	4544.848
PT.UNDERGRAD	1	131	472	1314	21836	1081.527	472	1672.203
IN.STATE.TUTION	480	2570	8050	11600	25750	7897.274	8050	5348.163
OUT.STATE.TUTION	1044	6108	8670	11660	25750	9276.906	8670	4170.771
ROOM.BOARD.COST	1260	3320	4030.5	4850	8700	4162.107	4030.5	1179.283
ROOM.COST	500	1710	2200	3040	7400	2514.682	2200	1150.837
BOARD.COST	531	1618.5	1980	2403	6250	2060.984	1980	661.7421
ADD.FEES	9	130	264.5	480	4374	392.0126	264.5	469.3792
BOOKS.COST	90	480	502	600	2340	549.9729	502	167.3554
PERS.SPENDING	75	900	1250	1794	6900	1389.292	1250	714.2479
PHD.FACT	8	57	71	82	105	68.64567	71	17.82563
TERM.DEG.FACT	20	63	77	90	100	75.23113	77	17.10816
STUDENT.FACT	2.3	11.8	14.3	17.6	91.8	14.85877	14.3	5.186399
ALUM.DONAT	0	11	19	29	81	20.91296	19	12.67414
EXPEND.PER.STUDENT	1834	6115.5	7729	10054	62469	8987.891	7729	5347.461
GRAD.RATE	8	47	60	74	118	60.40532	60	18.88906

Box plot of each variable:

Box plot of most of the variables indicate that there is high degree of variance due to presence of outliers. E.g. average number of applicants and average number of enrolled students varies across the colleges thus not normally distributed.





Variable Selection:

Before finding out linear relationship between predictors and response variable, we have to determine which predictor is highly correlated with response variable, graduation rate. But even after finding the set of predictor variable it is possible that with the set predictor variable could be correlated to each other i.e. possibility of multicolinearity.

Therefore, first I calculated correlation among predictor variables using "spearman" correlation rank algorithm and if correlation between any pair of variables is more than 0.85, discarded the second variable. From this I found out unrelated predictor variable. Between those predictor variables and response variable correlation was computed and "highly" correlated variable was chosen to do regression analysis.

Number of unrelated predictor variable selected using above method and their correlation matrix is given in next page. Using these variables again correlation was computed against response variable, graduation rate. This reduced number of variables to 6.

Number of variables highly correlated to response variable:

PUBLIC.PRIVATE	0.421
AVG.MATH.SAT	0.538
OUT.STATE.TUTION	0.633
ROOM.BOARD.COST	0.484
ALUM.DONAT	0.5
EXPEND.PER.STUDENT	0.485

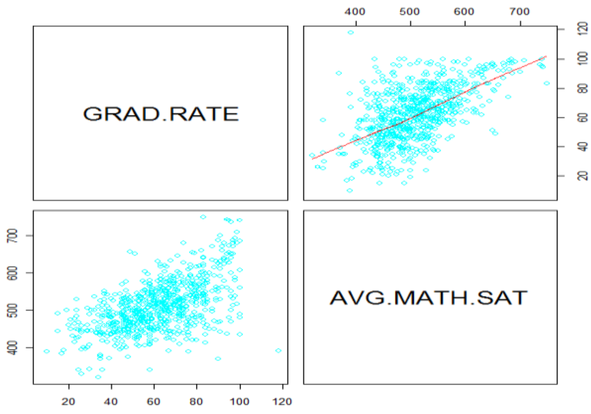
These are the variables which will be used to build linear regression model.

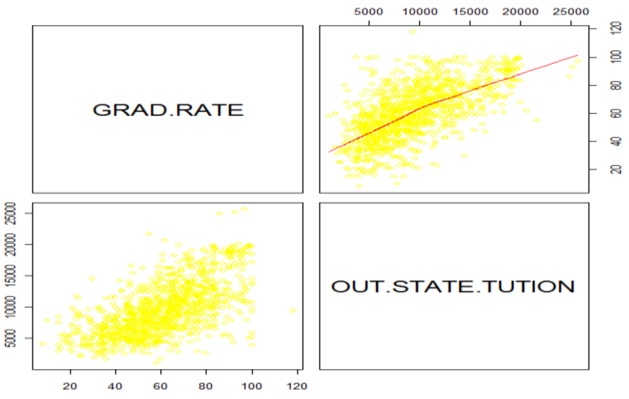
Univariate model:

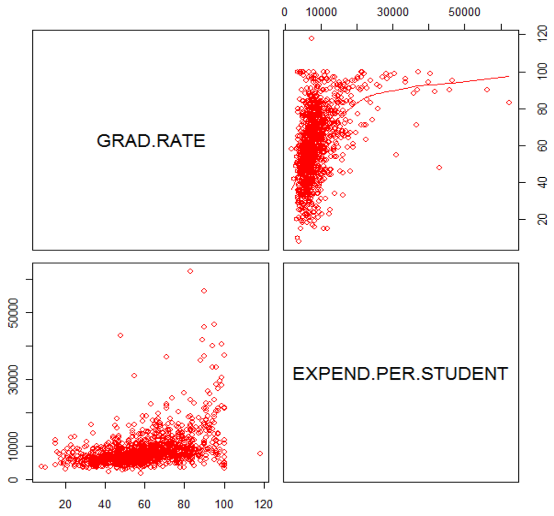
The regression and pairwise plot for graduation rate versus each of the above mentioned predictor variable are shown in next page. The plot indicates there is some degree of linear relationship exists except for public/private institution. One possible way of correcting this is by dividing data in two classes and constructing model for private and public institution separately. This is part of future work but for now I have not divide the data for doing further analysis.

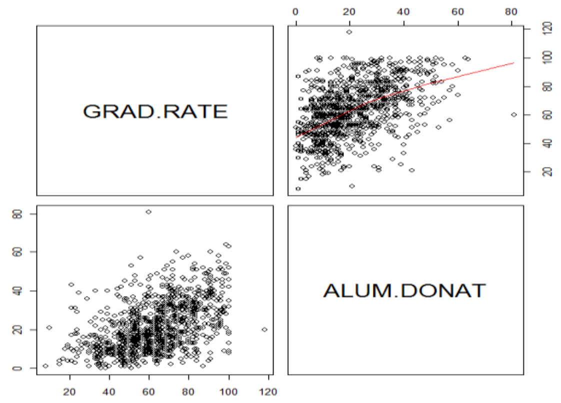
The most interesting feature of the plots is as predictor variable increases the rates of graduation increases e.g. increase in average math score increases the probability graduation.

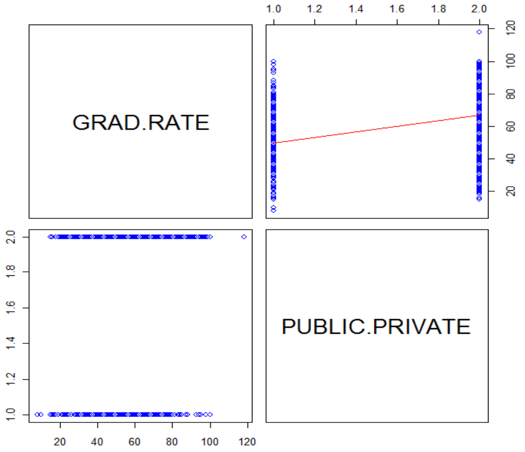
Pairwise plot and regression line:

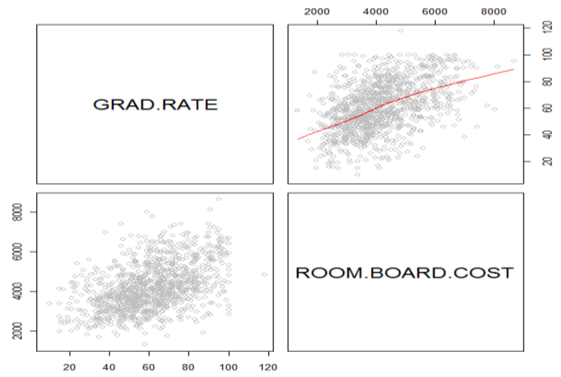








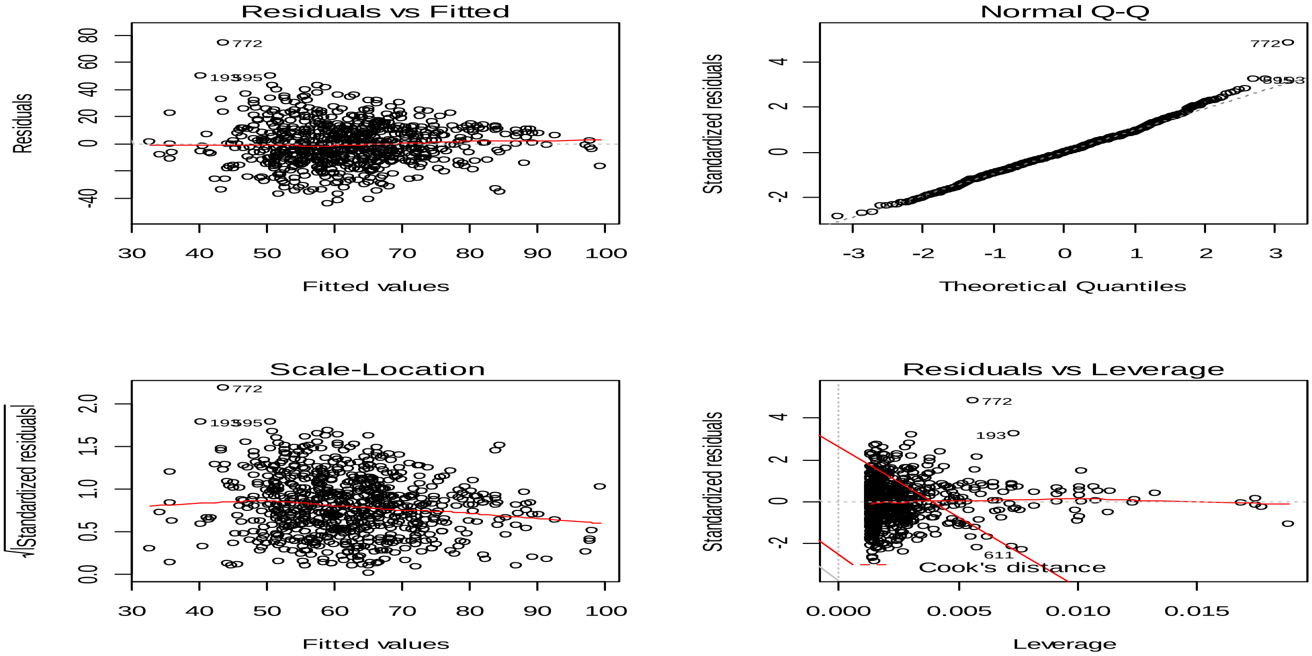




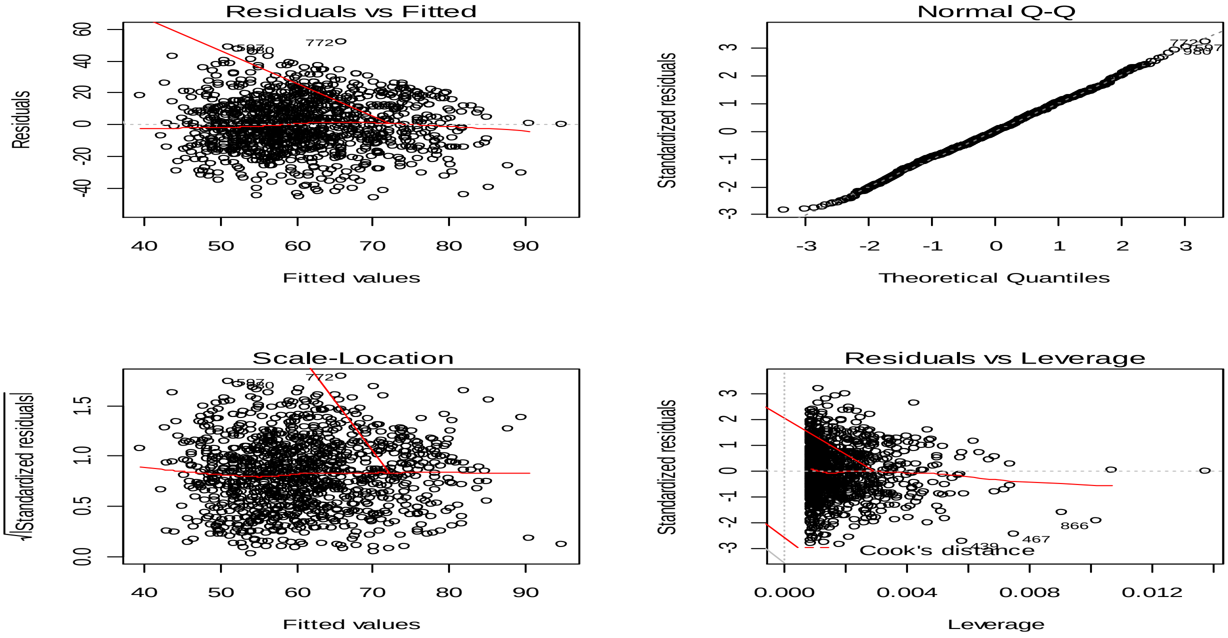
R squared value for each model is pretty low but that could be due to variance and possible outliers in the data. But model's constant variance assumption is valid, except for average SAT math score but the normality assumption is valid for each regression model. Model diagnostic plot for couple of variables is on next page and here is the summary of 6 model fit against response variable.

	R Squared	Spearman correlation	Chi-square	p-value
PUBLIC.PRIVATE	0.1067	0.421	2.632663	0.104686
AVG.MATH.SAT	0.2313	0.538	10.14075	0.001450275
OUT.STATE.TUTION	0.3026	0.633	2.937883	0.08652408
ROOM.BOARD.COST	0.131	0.484	0.9323532	0.3342522
ALUM.DONAT	0.2701	0.5	1.377788	0.2404782
EXPEND.PER.STUDENT	0.1144	0.485	2.41437	0.1202263

Out of state tuition plot:



Alumni donations:



Conclusion:

Based on above obtained result, I can say that there is weak relationship between individual predictor variable and graduation rate. Maybe transformation like boxcox/bulge rule or analyzing public/private universities separately or removing outliers may improve result.