COLLEGE GRADUATION RATES: DIAGNOSTIC MEASURES AND SUMMARY

LINEAR REGRESSION PROJECT:
FINAL REPORT

By

Parag Shah

INTRODUCTION

Problem Description and objective:

The importance that is placed on graduation rates as a measure of the success of institutions warrant the research into understanding the determinants of this factor. This study examines the role of different student characteristics like math/verbal SAT scores, personnel spending etc. and institutional factors like student to faculty ratio, number of PHD faculty, tuition etc. So 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.

PRELIMINARY ANALYSIS

The graduation rate of a university is the result of incorporating inputs from both the students and the institution. I am trying to find out determinants of graduation rates using multivariate regression analysis. Before doing the analysis selection of input variable was carried out. For this first correlation between input variables was found. Some of the variables were dropped if they are highly related with other variables (possible multicolinearity) and then final list of variables were determined based on correlation with response variable.

Correlation among some of the input variable were pretty high due to this around 15 variables were discarded and then rest of the variable were checked against response variable. Here correlation was not high so I kept threshold lower i.e. 0.4. This led to selection of 6 input variables. So for initial model 6 potential input variables were considered, public/private institution, average math SAT score, out of state tuition, room and boarding cost, alumni donation and student personnel expenditure. I tried to cover both institutional factors like public/private institution, out of state tuition cost, alumni donation and student characteristics like average math SAT score and personnel expenditure.

These individual input variables were regresses against response variable. The fit was poor, this could be because of presence of outliers in the data or data itself contained large variations among type of institution like public versus private. But model assumptions were plausible.

Summary of regression:

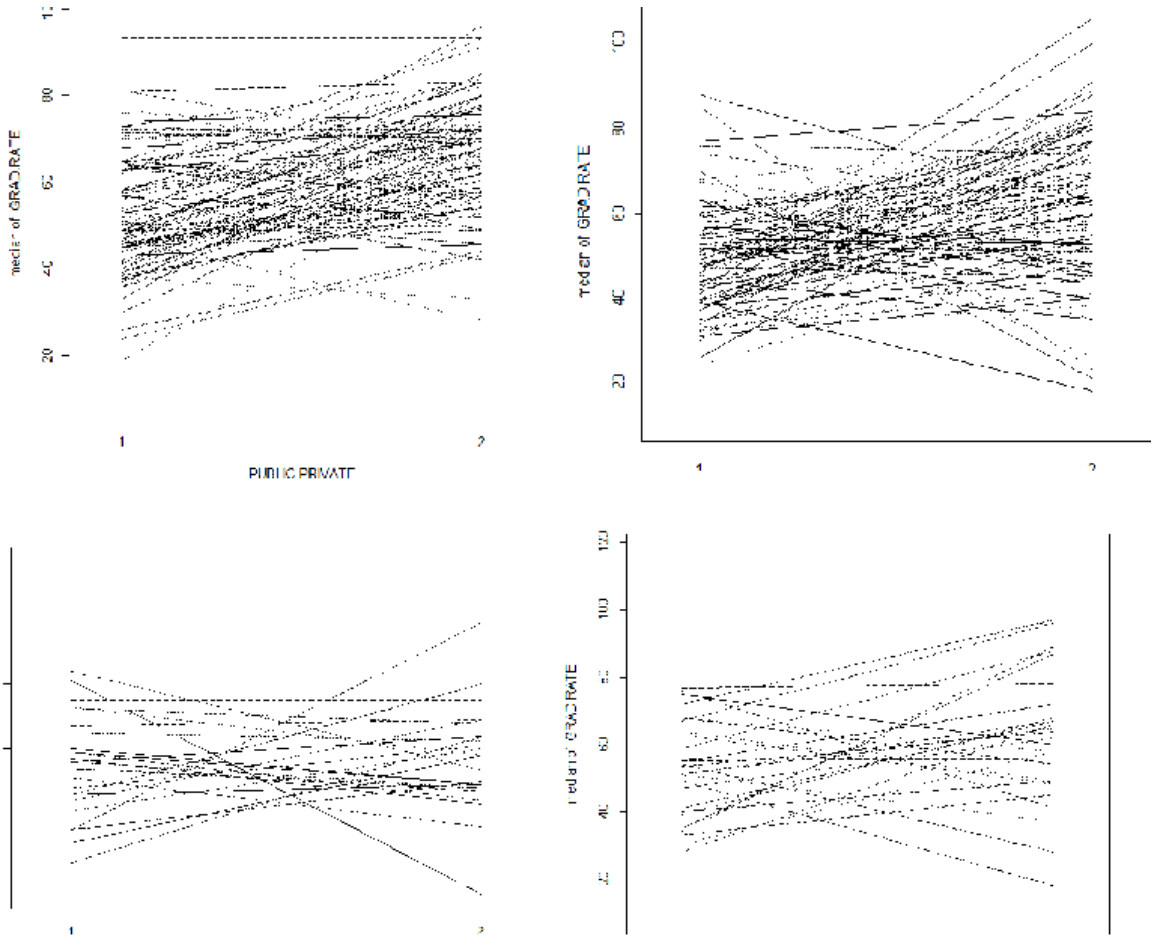
	R Squared	Spearman correlation	Chi-square	p-value
PUBLIC.PRIVATE	0.1067	0.421	2.632663	0.104686
AVG.MATH.SAT	0.2313	0.538	10.14075	0.001450275
OUT.STATE.TUTION	0.3026	0.633	2.937883	0.08652408
ROOM.BOARD.COST	0.131	0.484	0.9323532	0.3342522
ALUM.DONAT	0.2701	0.5	1.377788	0.2404782
EXPEND.PER.STUDENT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

MODEL CONSTRUCTION REFINEMENT

Using above chosen variables multiple regression was carried out. Again the fit was poor and some of the model assumptions were violated especially constant variance test.

R-squared value: 0.4476
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 18.59844 Df = 1 p = 1.613529e-05

Using interaction plot variables with significant interaction, public/private institution, was added in the regression model. And based pairwise of input and response variable higher order variable was considered. In order to determine the order of input variable simple *bulge rule* was considered. Since most of the input variables were increasing function of response variable, square of the variable was taken.



So resulting regression function with estimated coefficients are:

Graduation rate =
-1590 + 624 public/private college ^2
+ 0.0144 math sat score ^ 2
+ 0.0000136 out of state tuition ^2
+ 0.00000314 expenditure per student
+ public/private college * math sat score
+ public/private college * out of state tuition
+ public/private college * expenditure per student

Multiple R-squared: 0.4552, Adjusted R-squared: 0.4482
Chisquare = 0.7338091 Df = 1 p = 0.3916513

This function satisfies most of the model assumptions but again fit wasn't improved which indicates that poor fitting is not due to lack of transformation but could be due to other factors like outliers/missing data etc.

Furthermore all the input variables were significant and the variance inflation factor for each predicting variable was below 10, indicating that the problem of multicollinearity has been eliminated. Also the coefficients suggest that the variables are weighted slightly differently to obtain the graduation rates among colleges. The negative beta_0 doesn't tell anything about graduation rate since it should never go below 0.

MODEL VALIDATION

Another way to construct a multivariate model is by using automated search function to determine best possible candidate variables which can explain possible variation in response variable.

Using "step" function one set of variables identified which are:

> PUBLIC.PRIVATE, AVG.COMBINED.SAT, THIRD.QTL.MATH, FIRST.QTL.VERBAL, ENROLLED, PT.UNDERGRAD, OUT.STATE.TUTION, ROOM.BOARD.COST, ADD.FEES

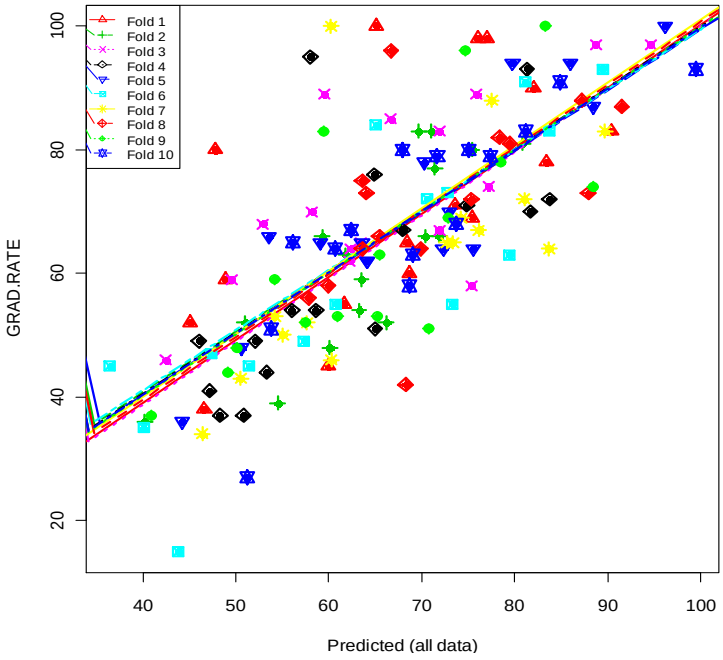
And using "regsubset" function another set of variables identified. Both of these subsets provided similar fit which was indicated by R-squared value but model assumption were violated by subset found using "regsubset". AIC score was also lower for this model.

Also one more model was constructed using "step" function model by adding interaction parameter for public/private institution. Its R-squared value and chi-squared value was improved but AIC value was not so decided to stick with original "step" function model.

Here is brief summary of all the models so far considered:

Models	**R-squared/Adj R-squared**	**Chi-squared value**	**Normality and other assumption satisfied?**	**AIC score**
Initial interaction model. (1)	0.4552 / 0.4482	0.7338091	Plausible	11127.4
Model found using "step" function. (2)	0.5391/0.5034	3.079416	Plausible	1229.865
Adding interaction terms in above model. (3)	0.5575/0.5094	2.335641	Plausible	1231.577
Model found using "regsubset" function. (4)	0.5433/0.4964	5.981774	No	1638.102

All the models were validated using 10-fold cross validation which also gave lower sum of squared values for model 2.



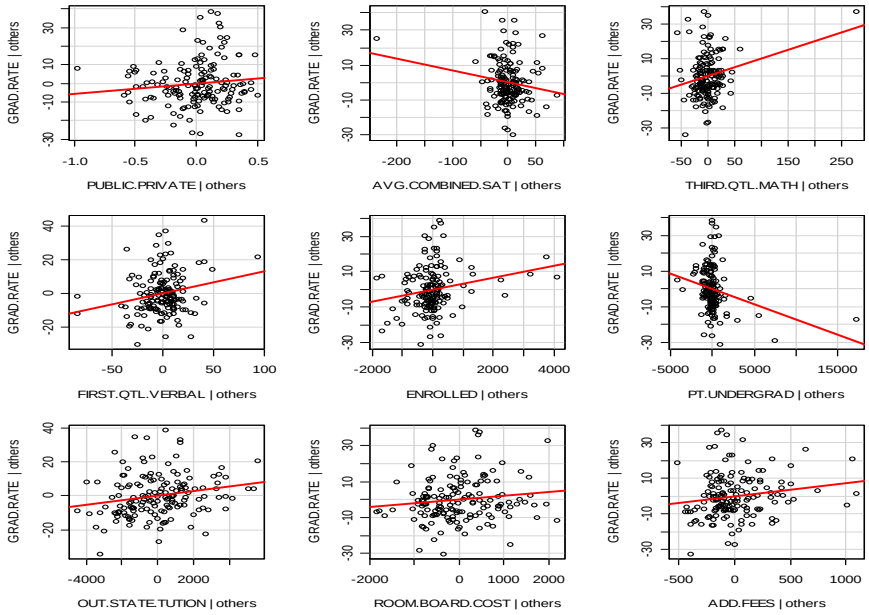
Using model 2 I can say that around 54% of variability in graduation rates can be explained by institution characteristics like type, number of students enrolled, number of part-time undergrad, out of state tuition, additional fees and room, boarding cost and by student characteristics like average combined SAT score, first quartile verbal, third quartile math.

Based on above findings I think predictive power and overall fit can be improved by removing the outliers and dividing the data by public versus private institution. This would also provide quality information about type of institution and their respective graduation rates. This can be taken further by combining this data with the rankings of institution and carrying further analysis.

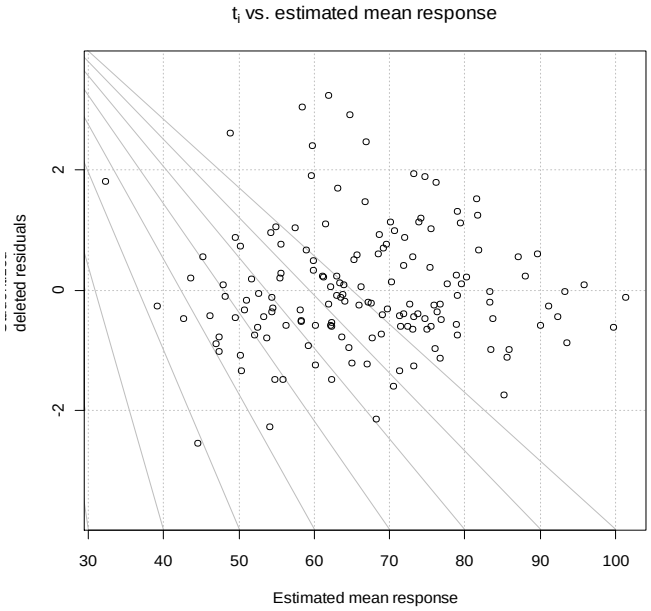
REGRESSION DIAGNOSTICS

Additional regression diagnostics were carried out using added variable plots, leverage plots. Also to detect outliers DFFITS, DFBETA procedure was used.

Marginal affect of adding a variable when other variable was already present was computed using added variable plot. From plot below it seems room and board cost should not be added to the model.



Studentized residual plot of fitted residuals which are under 95% confidence intervels.



Outlier analysis was carried using DFFITS and DFBETAS function using all the input variables. After removing the outliers fit was improved a bit i.e. R-squared was improved from 54% in model 2 to 61% but adjusted R-squared was reduced due to additional parameters. Also all the model assumption were validated.

Multiple R-squared: 0.605, Adjusted R-squared: 0.404
F-statistic: 3.01 on 31 and 61 DF, p-value: 0.000117

CONCLUSION

Even though analysis can be further carried out by dividing data set the resulting R-squared value obtained regression model has potential for predicting graduation rates. Maybe by including student behavioral factor (number of hours worked etc.) or demographic information fit can be improved and prediction can be more accurate with or without dividing data set.

References:

Dataset obtained from
http://lib.stat.cmu.edu/datasets/colleges/