COLLEGE GRADUATION RATES: REFINING MULTIPLE REGRESSION MODELS

LINEAR REGRESSION PROJECT: REPORT III

By

Parag Shah

Problem Description and objective:

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

Report Objective:

In the previous report multiple regression model was fitted using most correlated variable with graduation rate and in the process removed multicolinearity among variable by choosing one explanatory variable among list of highly correlated variables.

In this report I'll refine the model by using different set of dependent variables and validate model assumption, report possible multicolinearity and compare model to the model in the previous report.

New Model Selection criteria:

Model refinement and finding best possible explanatory variable was carried out using "regsubsets" and "step" method of R package "leaps". For finding best possible subset of explanatory variables all the dependent variable was passed to the function.

Note: I could not use simpler version of "leaps" function since number of dependent variables were more than 30.

Model selection using "step" function:

First all the dependent variable was fitted using regular multiple regression method (omitting missing values from data). And then ran step function to find best subset of possible explanatory variables. "step" method finds the best subset by adding and dropping variables and comparing the model based on AIC score. After several iterations it reported best model. Also on best model found, added the interaction terms and checked if it improves R squared or AIC values.

Residuals:

Min	1Q	Median	3Q	Max
-29.595	-7.211	-2.130	6.977	38.087

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.85E+01	1.55E+01	-1.195	0.233916	
PUBLIC.PRIVATE	5.40E+00	3.88E+00	1.39	0.166708	
AVG.COMBINED.SAT	-6.72E-02	3.43E-02	-1.957	0.05236	.
THIRD.QTL.MATH	1.02E-01	3.44E-02	2.956	0.003655	**
FIRST.QTL.VERBAL	1.32E-01	5.05E-02	2.606	0.010128	*
ENROLLED	3.33E-03	1.32E-03	2.533	0.012386	*
PT.UNDERGRAD	-1.72E-03	5.52E-04	-3.116	0.002219	**
OUT.STATE.TUTION	1.38E-03	5.27E-04	2.627	0.009555	**
ROOM.BOARD.COST	2.00E-03	1.32E-03	1.524	0.129835	
ADD.FEES	7.51E-03	3.75E-03	2.003	0.047103	*
ALUM.DONAT	4.14E-01	1.09E-01	3.796	0.000217	***
EXPEND.PER.STUDEN	-1.108e-03	2.96E-04	-3.742	0.000264	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.56 on 142 degrees of freedom
Multiple R-squared: 0.5391, Adjusted R-squared: 0.5034
F-statistic: 15.1 on 11 and 142 DF, p-value: < 2.2e-16

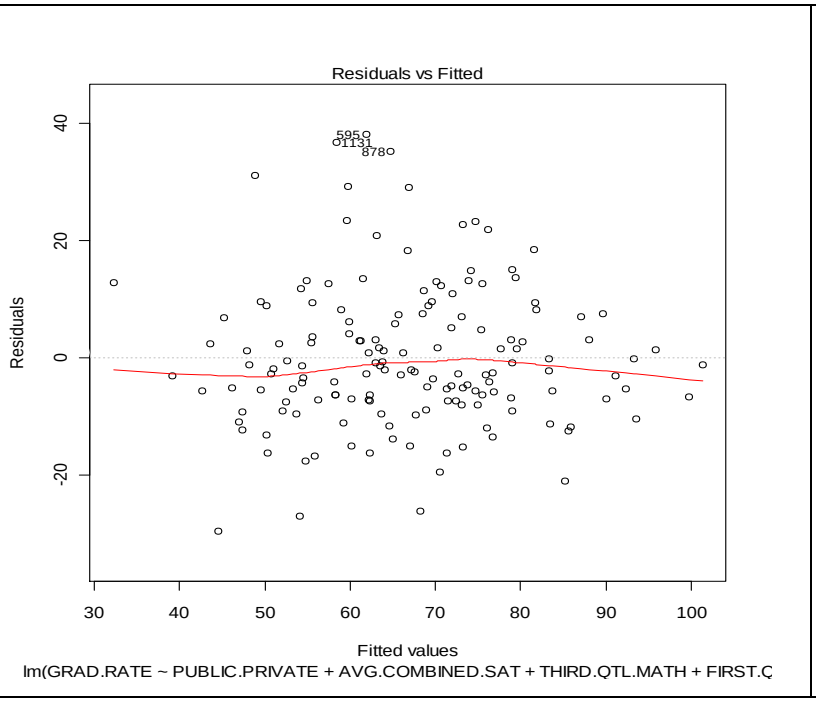
Negative value (which is very close to 0) of expenditure per student indicates that graduation rate tends to decrease as expenditure increases, same thing for number of part time undergrad.

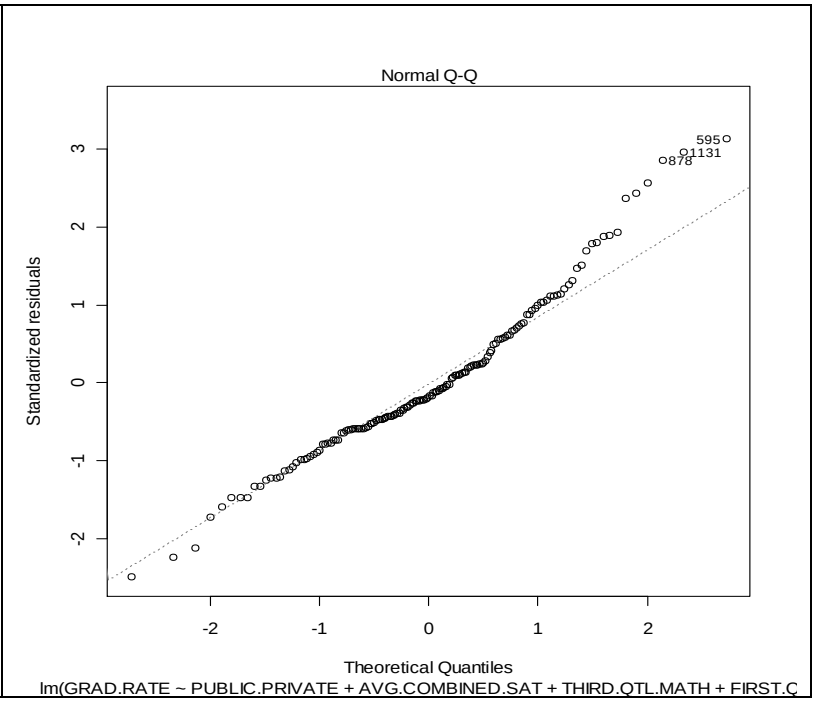
R-squared is much better than previous model. Also constant variance assumption was not violated.

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.079416 Df = 1 p = 0.07928899

Variance inflation factor for this models suggest possible multicolinearity associated with "average combined SAT" score. But rest of variable score is less than 10 which suggest colinearity is limited to only one variable. Although

residuals vs fitted value showed some deviation and thus some degree of over fitting but still it is not bad compared previous model. Also normality assumption is plausible.

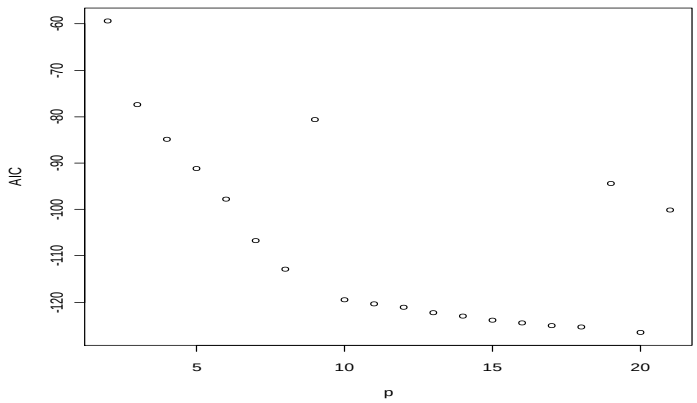




Model selection using "regsubset" function:

Same as above all the dependent variables were considered for finding best possible subset using "sequential replacement search" function. Regsubset function returns AIC score of each subset so minimum AIC associated with subset was chosen as best subset on which analysis was carried out.

Plot of AIC score:



Based on above value subset of 19 variables were selected. Although R-squared value was slightly higher than model presented above however assumption of constant variance was violated.

The variables which were considered:

GRAD.RATE ~ PUBLIC.PRIVATE + AVG.COMBINED.SAT
+ FIRST.QTL.MATH + THIRD.QTL.MATH + FIRST.QTL.VERBAL
+ FIRST.QTL.SAT + APPLICANTS + ACCEPTED + ENROLLED
+ PT.UNDERGRAD + OUT.STATE.TUTION + ROOM.COST + ADD.FEES
+ BOOKS.COST + PERS.SPENDING + PHD.FACT + TERM.DEG.FACT
+ ALUM.DONAT + EXPEND.PER.STUDENT

Residual standard error: 12.49 on 185 degrees of freedom
(1097 observations deleted due to missingness)
Multiple R-squared: 0.5433, **Adjusted R-squared: 0.4964**
F-statistic: 11.58 on 19 and 185 DF, p-value: < 2.2e-16

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 5.981774 Df = 1 p = 0.01445446

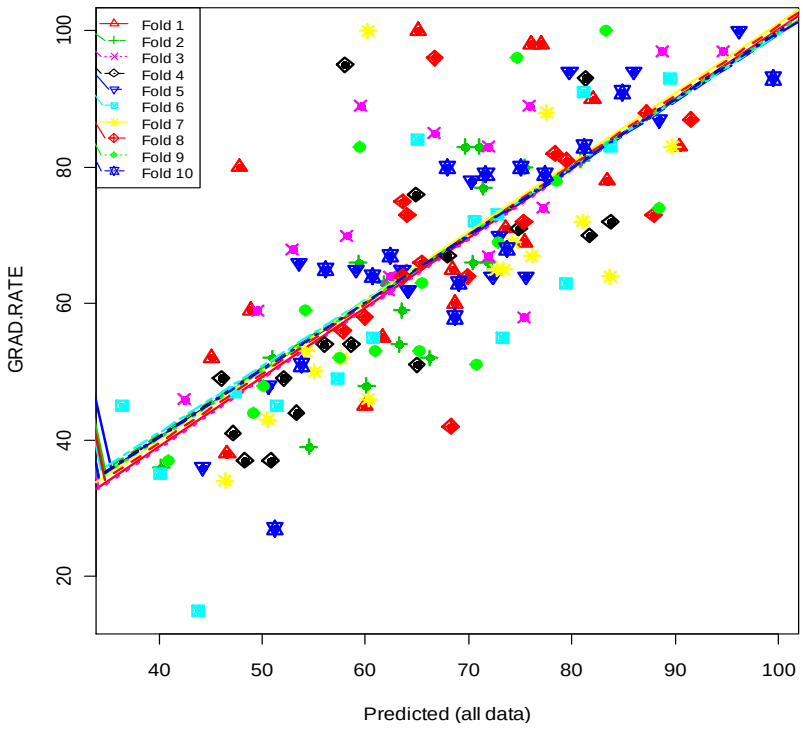
Summary of 4 best models so far:

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

Based on 4 criteria I think best model is model 2 since it has higher R-squared value, plausible model assumption, gives constant variance and lower AIC value compared to other two.

Model Validation using k-fold cross validation:

On above model 10-fold cross validation was performed to check model's predictive power.



Conclusion:

Graduation rate among colleges can be explained by public vs private college, average combined SAT, third and first quartile SAT, number of students enrolled, number of part time undergrad, out of state tuition, room cost, additional fees, alumni donation and expenditure per student. An estimated 54% of variability is accounted for by explanatory variables in model 2 and their associated regression coefficients. Also its predictive power and associated AIC score validates that model is better than other 3.