COLLEGE GRADUATION RATES: MULTIPLE REGRESSION MODELS

LINEAR REGRESSION PROJECT: REPORT II

By

Parag Shah

MULTIVARIATE ANALYSIS

After analyzing the data we found 14 variables which were most related to response variable i.e. graduation rate. And after detecting colinearity we reduced number of input variable to 6. Which are:

	Spearman correlation
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

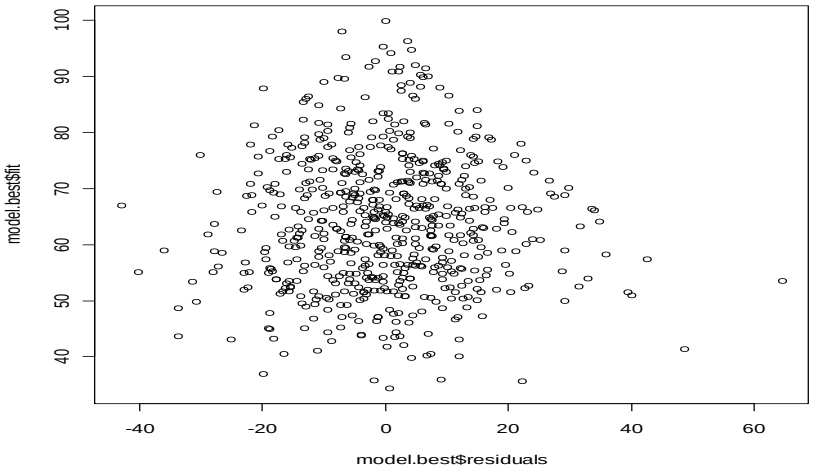
To assess the linear relationship with respect to response variable formal test were done. T and F test were done and based on 95% confidence interval it was concluded that all the variables are significant and none of them are equal to 0.

So I fitted multiple regression models using these variables. The R-squared value was not that high: 0.4476 but major concerns was variability in the model. To test this chi-square (breusch-pagan) test was done.

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

High Chi-square value and low p value suggest rejection of null-hypothesis (which is homoscedasticity). Assumption of constant variance was violated. This is not surprising as box plots indicate that larger spread of data for the overall graduation rates of colleges and other input variables.

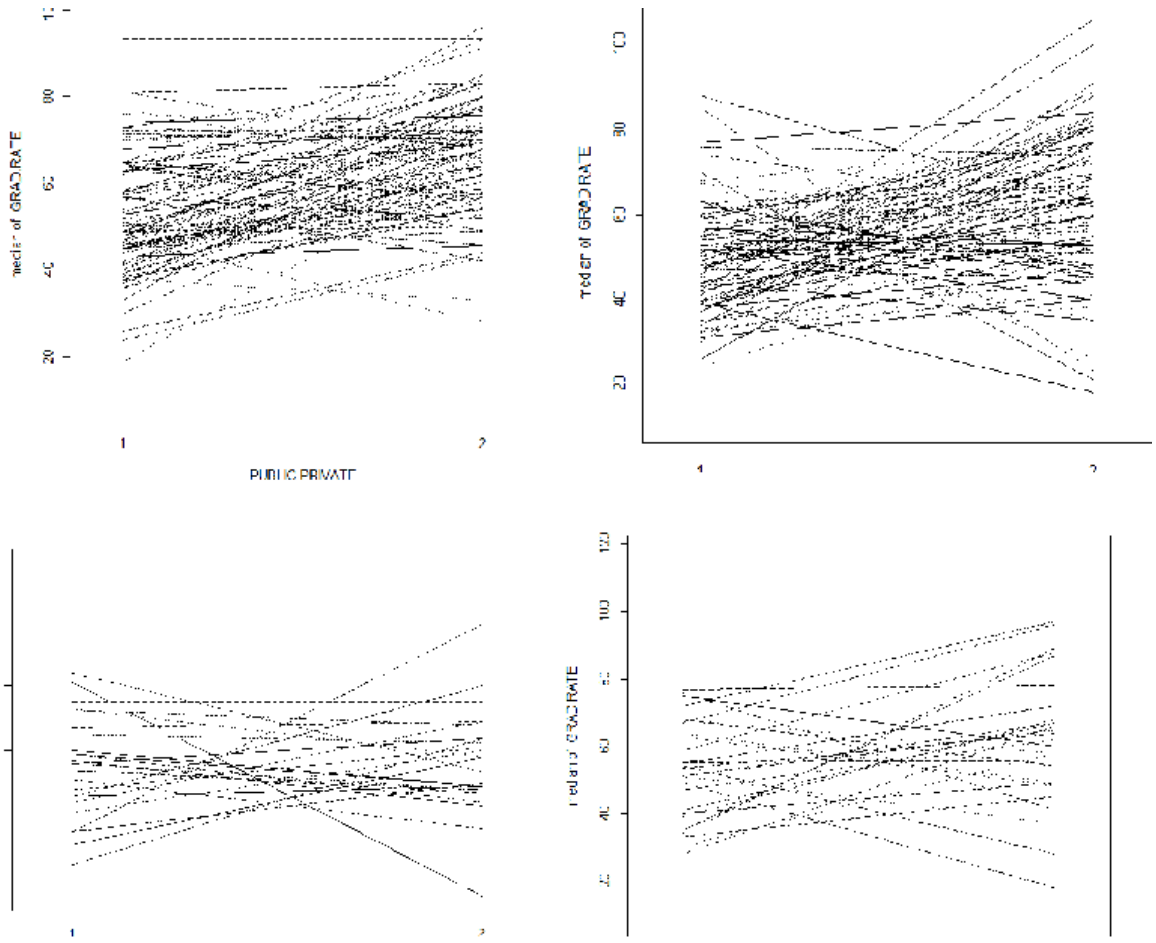
Residual versus fitted data plot for model with 6 variables:



In order to correct the heteroscedasticity, number of different model was considered. A log-log model was considered. Logging the individual input variables, however, did not provide the desired results. The data actually became more skewed after the transformation. Attempts to build a model by logging all the predictor variables or just the few variables were not successful in correcting the non-constant variance problem.

Based on interaction plot of individual variables, interaction model was considered. With Public/private school and average of SAT math, out of state tuition and expense per student were added as interaction terms.

Interaction terms prove to give more satisfactory result in terms of getting homoscedasticity.



Below are the predictor variables and its coefficients obtained by adding interaction terms in linear regression model.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.59E+03	4.03E+02	-3.935	9.26E-05	***
I(PUBLIC.PRIVATE^2)	6.24E+02	1.97E+02	3.172	0.00159	**
I(AVG.MATH.SAT^2)	1.44E-02	1.43E-03	10.1	< 2e-16	***
I(OUT.STATE.TUTION^2)	1.36E-05	4.90E-06	2.784	0.005542	**
I(EXPEND.PER.STUDENT^2)	3.14E-06	8.72E-07	3.598	0.000347	***
PUBLIC.PRIVATE:ROOM.BOARD.COST	2.20E-01	4.62E-02	4.767	2.34E-06	***
PUBLIC.PRIVATE:OUT.STATE.TUTION	-1.17E-01	6.28E-02	-1.87E+00	0.06232	.
PUBLIC.PRIVATE:ALUM.DONAT	1.75E+01	3.58E+00	4.88	1.35E-06	***
PUBLIC.PRIVATE:EXPEND.PER.STUDENT	-1.25E-01	2.58E-02	-4.825	1.77E-06	***

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

Residuals:

Min	1Q	Median	3Q	Max
-4011.1	-1142.9	-116.4	940.2	10702.0

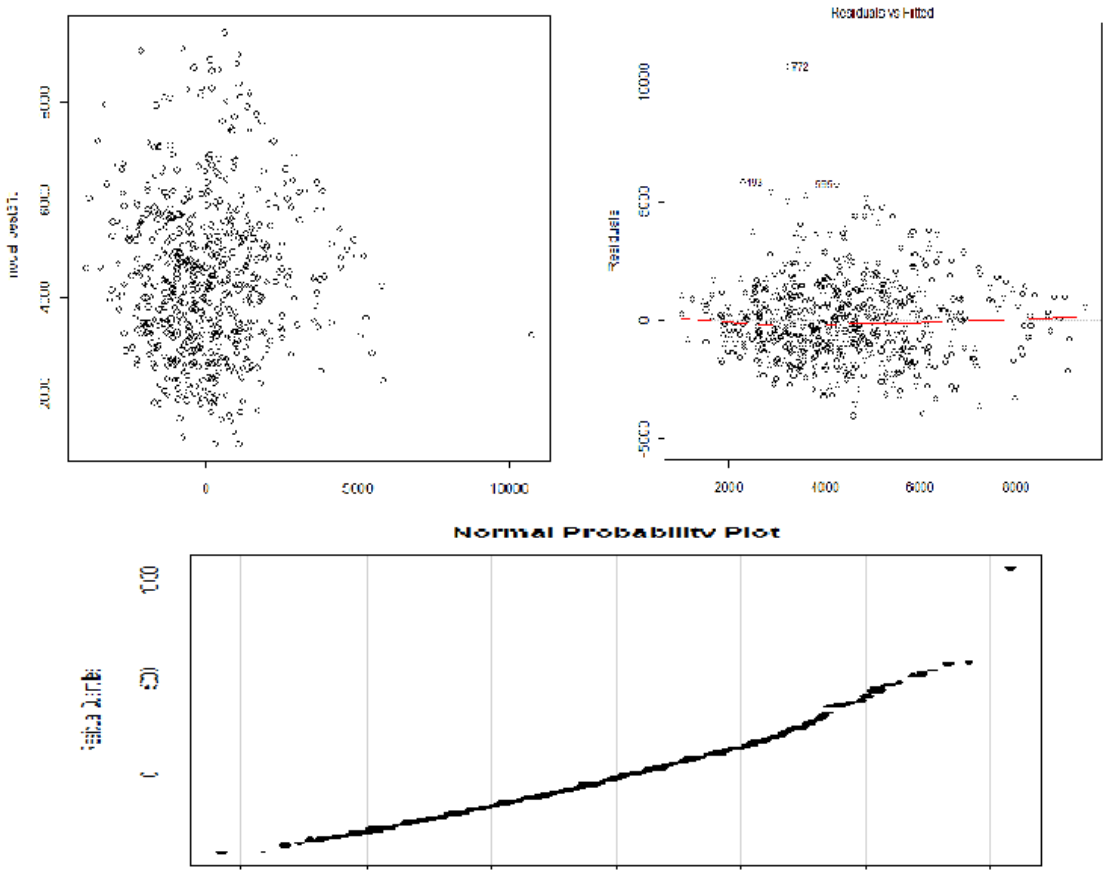
Residual standard error: 1704 on 618 degrees of freedom
(675 observations deleted due to missingness)
Multiple R-squared: 0.4552, Adjusted R-squared: 0.4482
F-statistic: 64.55 on 8 and 618 DF, p-value: < 2.2e-16

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.7338091 Df = 1 p = 0.3916513

All the input variables are significant and the variance inflation factor for each predicting variable was below 10, indicating that the problem of multicollinearity has been eliminated.

After adding interaction terms and fitting model all assumption including constant variance and assumption of normality was plausible.

Below are diagnostic plots for residuals versus fitted value and normal probability plot.



The same model was used on modified data i.e. data without any outliers. But there was no improvement in R-squared value and non-constant variance assumption was again violated. Outliers were removed using "outlier" package function in R. All outliers were removed based on median value of data. R-squared and chi-squared obtained was:

Multiple R-squared: 0.3012, Adjusted R-squared: 0.2973
Chisquare = 12.8832091

In addition to multiplicative interaction model on most correlated variable (after detecting colinearity) other methods were used to find significant input variable using "regsubsets" function on 31 input variables and graduation rate as response variable. Variables were chosen based on AIC score. Using those variables model was created which resulted in slightly higher R-squared value but increased variance.

Conclusion:

Since model with interaction terms keeps variance relatively constant, I think it can explain, with more precision, some of the factors affecting graduation rate in colleges.

The regression equation is given by:

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

The coefficients suggest that the variables are weighted slightly differently to obtain the graduation rates among colleges.

From the score, highest weight is given to public/private colleges. Also looking at the data it seems like public versus private colleges graduation rates significantly differ.

Note on R-squared value:

Although R squared value 0.48 which in general considered being low but I think for problem like this it is good enough since there is high variability among data points.