CSC 423 - ASSIGNMENT 5

BY

Sasidhar Mukthinuthalapati

**PROBLEM - 1**

1. Distribution of GradRate can be studied using a Histogram. To generate the histogram for the GradRate we run the following code in SAS:

title "Univariate analysis on GradRate";

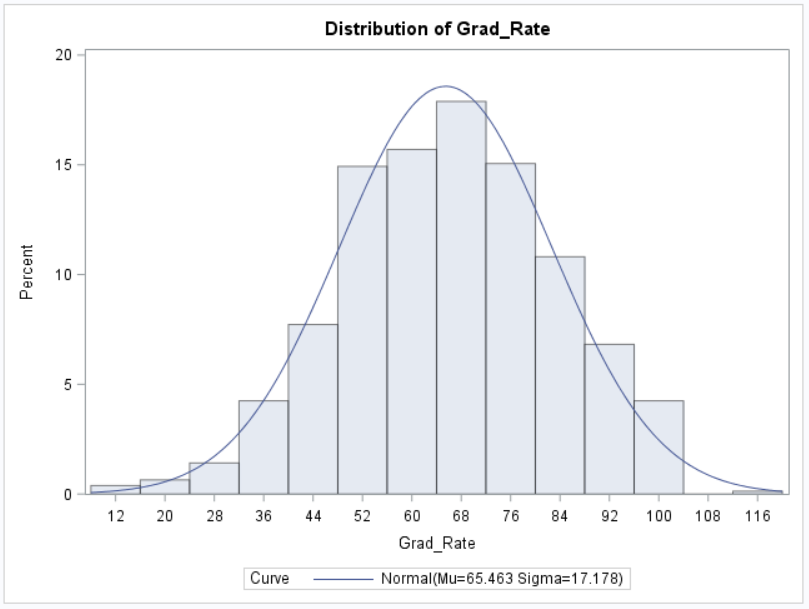
**proc** **univariate** normal;

var Grad\_Rate;

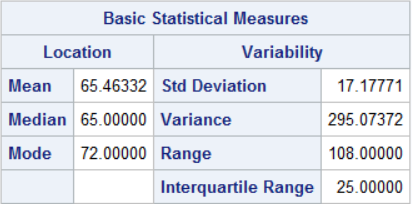
histogram / normal (mu = est sigma = est);

**run**;

The following output is generated:



As we can see from the above Histogram, it looks symmetric and there is no need of any kind of transformation.



As we can see from the above table that **Mean > Median** which means that the GradRate is **Positively Skewed** or **Right Skewed.**

IQR = 25.00, Q1 = 53 and Q3 = 78, So to check for outliers we compute the following Range (Q1 – 1.5\*IQR, Q3 + 1.5\*IQR) => (53 – 1.5\*25, 78 + 1.5\*25) => **(15.5, 115.5)** so any data points above or below this range will be considered as outliers and we can see that there are a few outliers in the GradRate distribution.



The Kurtosis and Skewness values are shown in the above picture. The skewness value is negative which tells is that the **left tail is longer than the right tail**.

The Kurtosis value is negative and closer to zero which means the **peak of the distribution is not so peaky or tall** which is evident from the Histogram.

1. To create the scatterplot matrix, we will run the following piece of code in SAS:

title "Scatterplot Matrix";

**proc** **sgscatter**;

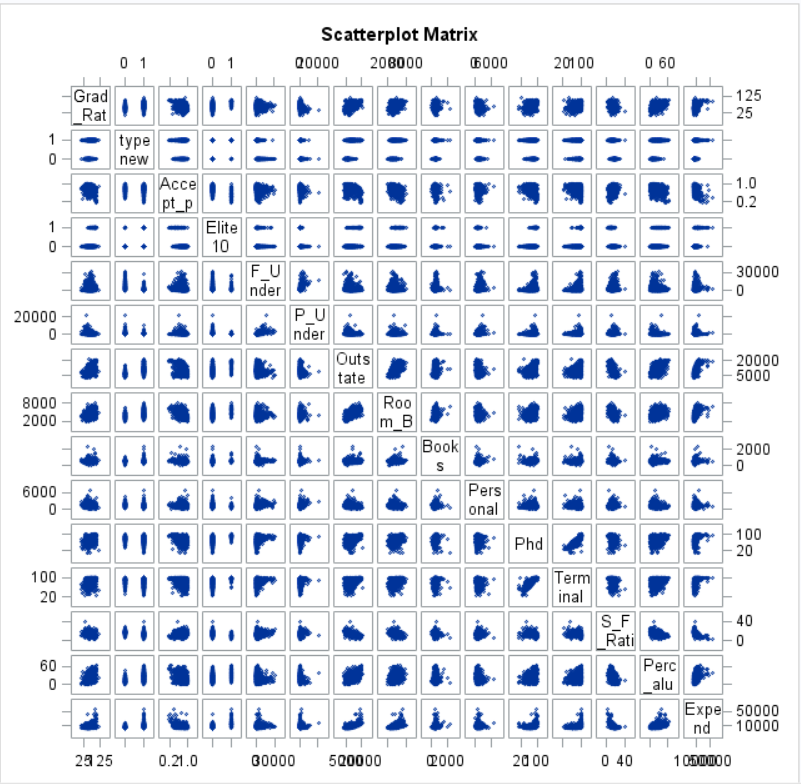
MATRIX Grad\_Rate typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend;

**run**;

From the scatter plots which were generated we can make the following observations:

* Few of the predictors in the complete model look to have linear relation but the slope of the line is very close to 90 Degrees.
* Grad\_Rate and Outstate seem to be poisitively and Linearly correlated to each other but the strength of the correlation cannot be determined using the scatterplot.
* Grad\_Rate and PhD seem to be positively and Linearly correlated to each other but the strength cannot be determined using the scatterplot.
* Grad\_Rate and Perc\_Alumni is also linear but can’t say.
* Most of the scatterplot of variables with respect to Grad\_Rate seem to be Random and hence much cannot be interpreted using this.

The Scatterplot Matrix is given below:



For the better understanding of the correlation of GradRate with respect to the other variables we will compute the Pearson’s Correlation value to determine which have strong correlation with GradRate.

To generate the correlation values we run the following code in SAS:

title "Correlation values";

**proc** **corr**;

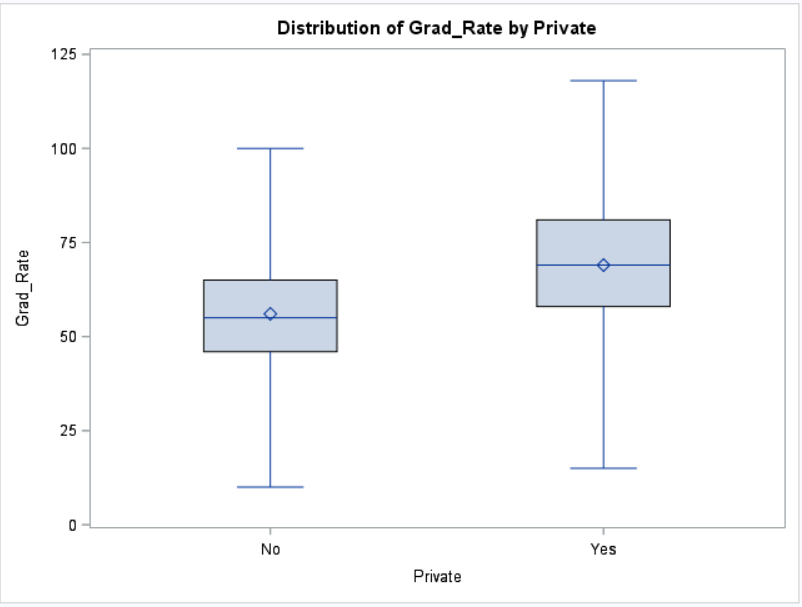
var Grad\_Rate typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend;

**run**;

We get the correlation values and we can make the following observations:

* GradRate is positively correlated to typenew (Dummy Variable for Private or Public), Elite10 ,Outstate, Room\_Board, Books, PhD, Terminal, Perc\_alumni and Expend.
* The remaining predictors lke S\_F\_Ratio, Personal, P\_Undergrad, F\_Undergrad, Accept\_Pct are negatively correlated to GradRate.
* The predictor which has the strongest correlation to GradRate among all the other variables is OutState.

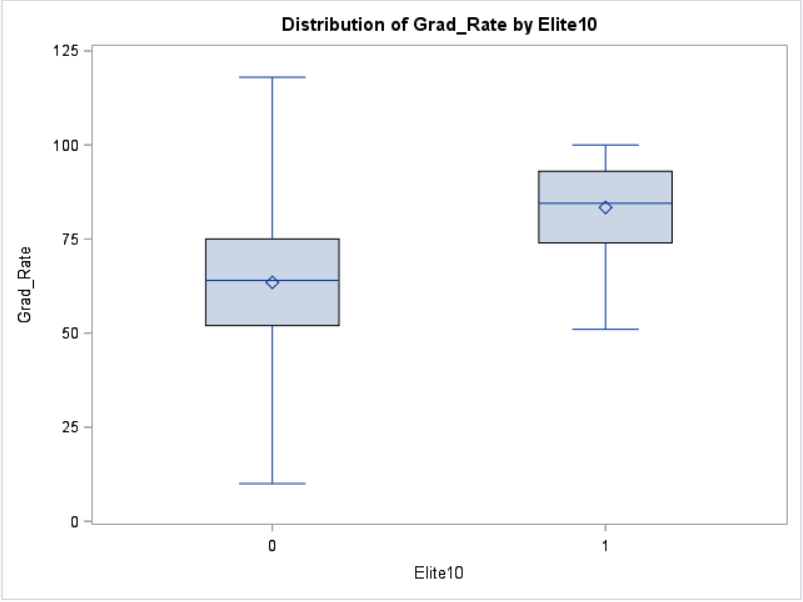
1. Boxplots of GradRate Vs Public or Private is as follows:



From the above boxplot, we can make the following observations:

* Mean GradRate is different in both Public and Private Universities. Private has higher Mean Grad\_Rate.
* We can also see that there might be outliers in the Private Universities dataset because the GradRate surely crosses 100% for few universities.
* We can also see that the mean and median are different for Public Universities which means that the Public Universities dataset might be skewed.
* Public Universities have the least minimum graduation rate when compared with private universities.

The Boxplot of GradRate vs Elite is as follows:



We can make the following observations from the above boxplot:

* The universities which have the elite students (1) have a higher mean Graduation rate when compared to the other class (0).
* The Mean and Median in both the class distribution is different which means that the distribution is skewed for both the classes.
* The Graduation Rate for the universities which have elite10 variable as 0 has few values which are greater than 100% which means there are outliers present.
* The universities in which the elite10 are present has a minimum graduation rate of nearly ~50%.

The code to generate the above-mentioned output is as follows:

title "BoxPlot of GradRate vs Public/Private ";

**proc** **sort**;

by Private;

**run**;

**proc** **boxplot**;

plot Grad\_Rate\*Private;

**run**;

title "BoxPlot of GradRate vs Elite";

**proc** **sort**;

by Elite10;

**run**;

**proc** **boxplot**;

plot Grad\_Rate\*Elite10;

**run**;

1. The code to generate the Full model which will predict the GradRate is as follows:

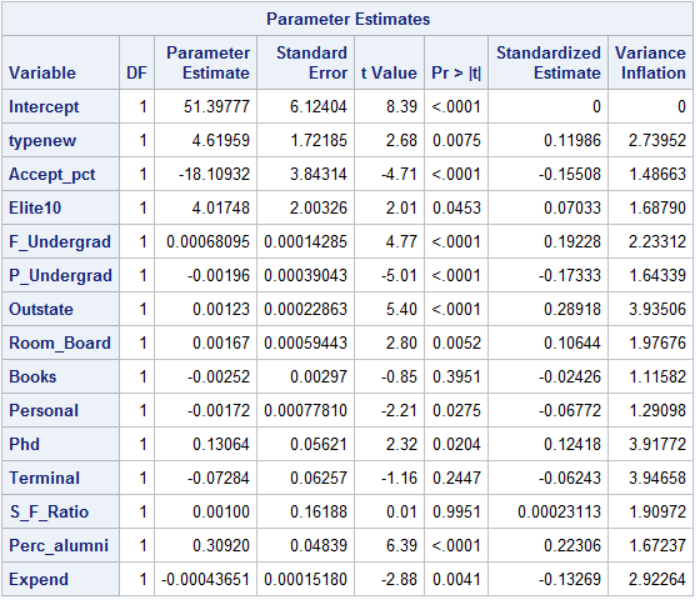
title "Full Model";

**proc** **reg**;

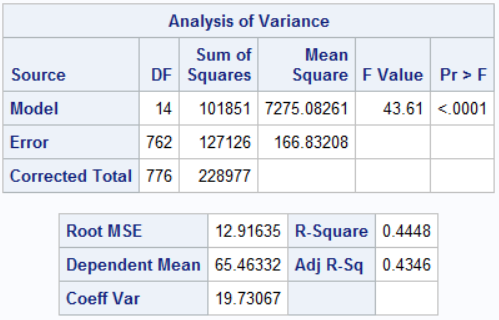
model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend / vif stb;

**run**;

The output and parameter estimates which are generated are as follows:



As we can see, the P-Value associated with Books, Terminal and S\_F\_Ratio is greater than 0.05 which tells us that these variables are in significant while computing the Graduate Rate and can be omitted from future models.



But for this full model as we can see that the **F-Value is 43.61 and the P-Value associated with it is <0.0001 which means the model is a good model**.

The **Adjusted R-Squared value tells us that 43.46% of the variance in Graduation Rate** can be explained using this model and the full list of predictors.

From the Parameter Estimates we can make the following observations:

* Typenew which is a dummy variable for the university type i.e. Private or Public and the base case is 0 when the university is Public and 1 when it is Private. The change of the dummy variable from 0 to 1 will change the graduation rate by 4.61% increase.
* 1 unit positive change in Acceptance Percentage would result the graduation rate to decrease by 18.10%.
* Change of Elite10 from 0 to 1 would result an increase in the graduation rate by 4.01%.
* 1 unit increase in the number of full time undergraduate student would result in the increase in the graduation rate by 0.00068%.
* 1 unit increase in the number of part time undergraduate students would result in the decrease in the graduation rate by 0.0019%.
* 1 Dollar increase in the out-state tuition would result in the increase in the graduation rate by 0.0012%.
* 1$ increase in the Room and Boarding costs would result in the increase in the graduation rate by 0.0016%.
* 1$ increase in the book cost would result in the graduation rate decreasing by 0.0025%.
* 1$ increase in the Personal spending would result in the decrease in the graduation rate by 0.0017%.
* 1 unit increase in the number of faculty having PhD would increase the graduation rate by 0.130%.
* 1 unit increase in the number of faculty with terminal degrees would decrease the graduation rate by 0.072%.
* 1 unit increase in the Student/Faculty Ratio would result in the increase in the graduation rate by 0.001%.
* 1% increase in the Percent of alumni who donate towards the university would result in the increase in the graduation rate by 0.309%.
* 1$ increase in the instructional expenditure per student would result in a negative change i.e. decrease in the graduation rate by 0.00043%.

The Complete Regression Equation for the Full Model is as follows:

**Grad\_Rate = 51.397 + 4.619\*Typenew – 18.10\*Accept\_pct + 4.01\*Elite10 + 0.0006\*F\_Undergrad – 0.0019\*P\_Undergrad + 0.0012\*Outstate + 0.0016\*Room\_Board – 0.00252\*Books – 0.0017\*Personal +0.13\*PhD - 0.072\*Terminal + 0.001\*S\_F\_Ratio + 0.309\*Perc\_Alumni – 0.0004\*Expend**

1. Multi-Collinearity doesn’t seem to be a problem in this case because when we compute the VIF Value for the predictors, none of the predictors have VIF Value greater than 10 which tells us that Multi Collinearity is not a problem in this dataset.

This can also been seen from the Scatterplot Matrix and the Pearson’s Correlation values.

The code to generate the Variance Influence values for each predictor is as follows:

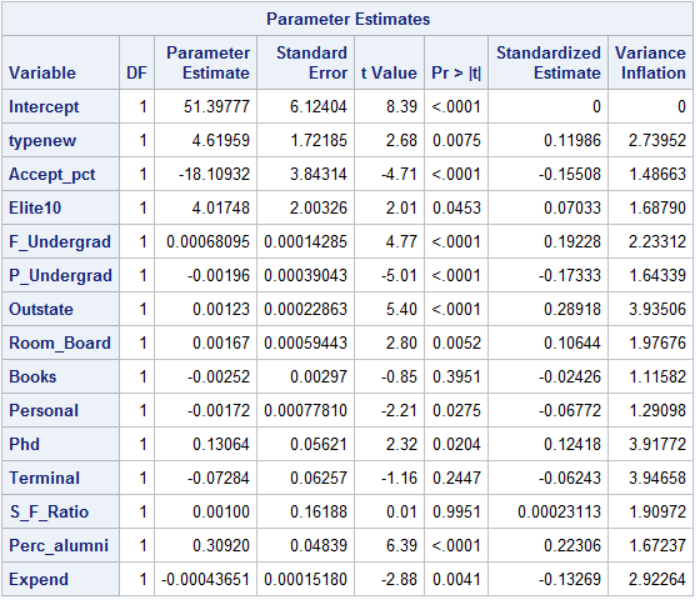
title "Full Model";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend / vif stb;

**run**;

The above code generates the following tables which has the VIF values:



1. For this question, I’ve chosen to go with Adjusted R-Squared and Forward Stepwise Selection:

First, I’ve run the stepwise forward selection method. In this method, SAS keeps adding one variable after the other.

To generate the output for this model we run the following piece of code:

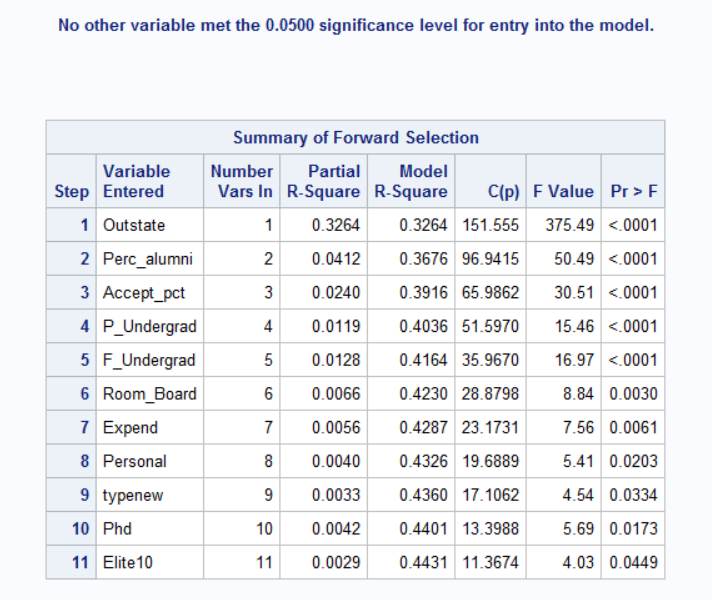
title "New model using Forward selection method";

**proc** **reg**;

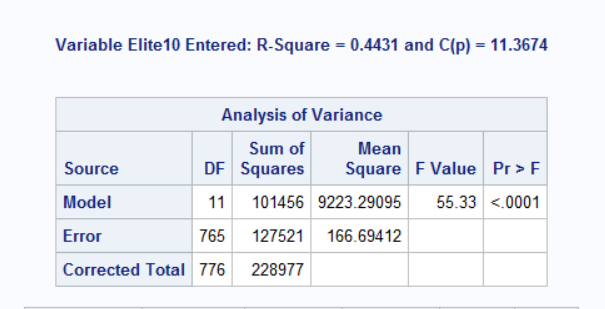
model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend / selection = forward sle = **0.05** sls = **0.05**;

**run**;

sls and sle are used to check the significance of an attribute on each step, so if a variable has significance which is less than greater than 0.05 then that variable will be removed from the upcoming models.



As we can see in this model all the variables have significance which is less than 0.05.



As we can see that the F-value and the P-value associated with this F-value is <0.0001 which means that this model is a good model.

The R-Squared value for this model is 0.4431 which means that nearly 44.31% of the variance in Graduation Rate could be explained using this model.

The other selection method is the Adjusted R-Squared method and the code to generate the models with their respective Adjusted R-Squared values is as follows:

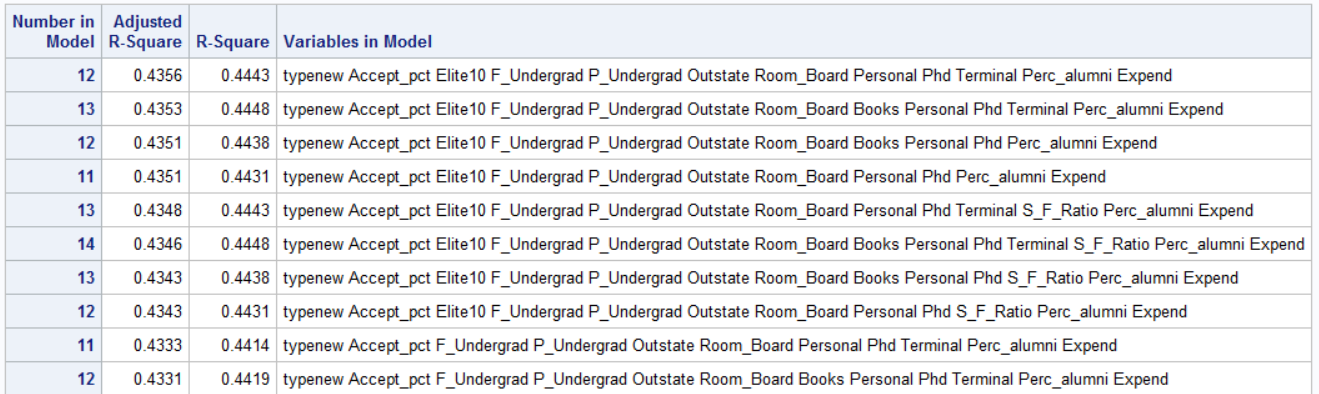
title "New model uaing adjusted R2 selection method";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend / selection = adjrsq sle = **0.05** sls = **0.05**;

**run**;

We get the following output:



And the table goes on but I’ve included only the top rows of the output. As we can see the model with the highest Adjusted R-Squared values are placed at the top.

Here it also includes the models which have insignificant variables in them. Like for example the first and second model in the table still have the insignificant variables in them.

So, based on **Adjusted R-Squared I would choose the 4th model in the above table which has 11 predictors** and has the adjusted R-Squared value which is very close to the best Adjusted R-squared value.

We can also see that in the model with the highest Adjusted R-Squared value, the variables Books and Student to Faculty Ratio were not used. This was also evident from the Significance test which was done as part of the Parameter Estimates.

1. I’ve chosen to go with the model which has **Adjusted R-Squared value = 0.4351 and has only 11 predictors**. In the previous question the model with the best Adjusted R-Squared value was the first model in the table but this model still had predictors which weren’t so significant.

The full expression for this model is as follows:

**Grad\_Rate = 48.4030 + 4.77\*typenew – 17.782\*Accept\_pct + 4.021\*Elite10 + 0.0006\*F\_Undergrad – 0.0019\*P\_Undergrad + 0.0012\*Outstate + 0.0015\*Room\_Board – 0.0018\*Personal + 0.084\*PhD + 0.30\*Perc\_alumni – 0.00044\*Expend**

In this model Books, Terminal and S\_F\_Ratio were not used because they weren’t significant which could be determined by the p-values associated with them.

The code to generate the above model is as follows:

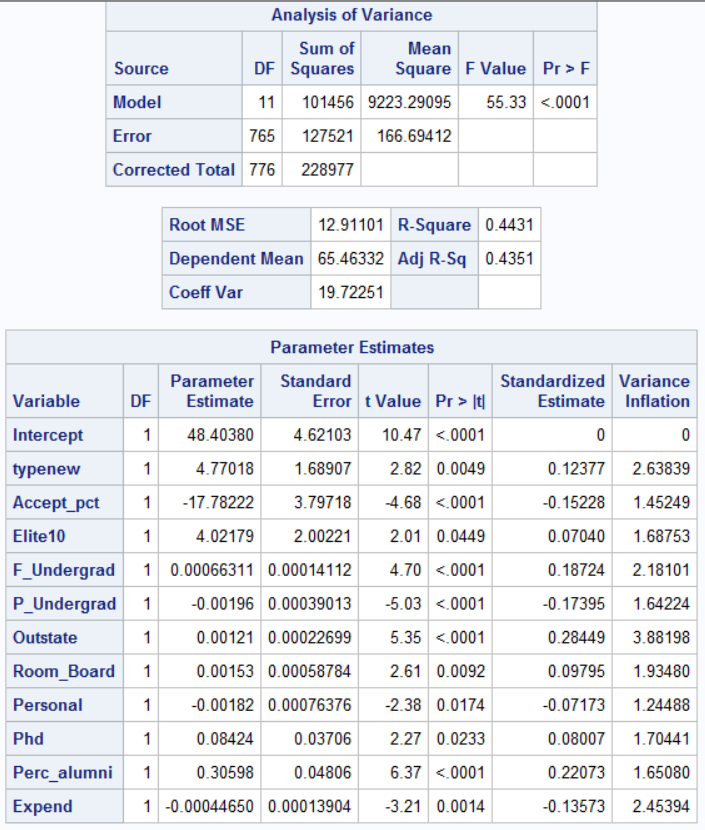
title "Final Model";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend / vif stb;

**run**;

The output generated by the above code is as follows:



From the above tables, we can see that the current model doesn’t have any insignificant attributes.

The Adjusted R-Squared value is **0.4351** which means that **43.51% of the variance in the Graduation Rate could be explained using this model**.

**This model is not having any multicollinearity problem** as the Variance Inflation values for all the attributes is less than 10.

The **F-value is 55.33 and the P-value associated with the F-Value is <0.0001** which means this is a good model.

1. Studentised vs Predicted Plot: To generate the plot we run the following piece of code:

title "Studentised values vs Predicted values Plot";

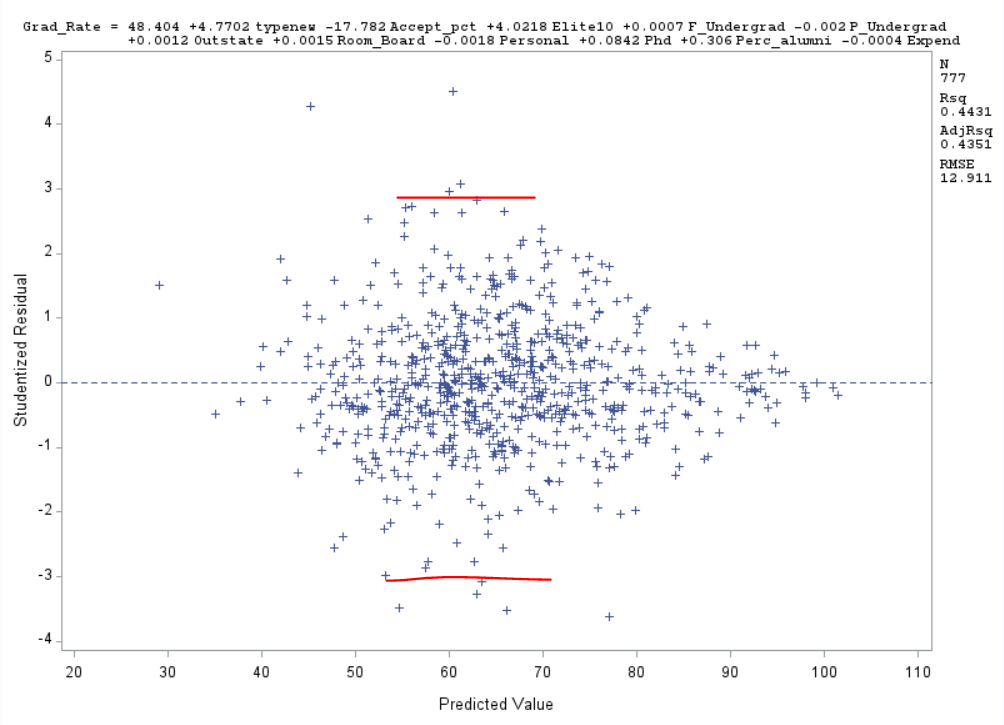
**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend;

plot student.\*predicted.;

**run**;

We get the following output:



We can make the following observations from the above plot:

* The dataset has quite a lot of outliers present in them. The points above and below the red line can be considered as outliers because their studentised residual values is greater than +/-3.
* The variance doesn’t seem constant throughout the plot.
* But we can see that the points are somewhat equally distributed above and below the 0-residual line.
* The points are kind of independent of each other.
* There is no striking pattern as such but we can see that most of the points are concentrated towards the center of the plot.

1. To generate the Normality plot we run the following piece of code:

title "Normality Plot";

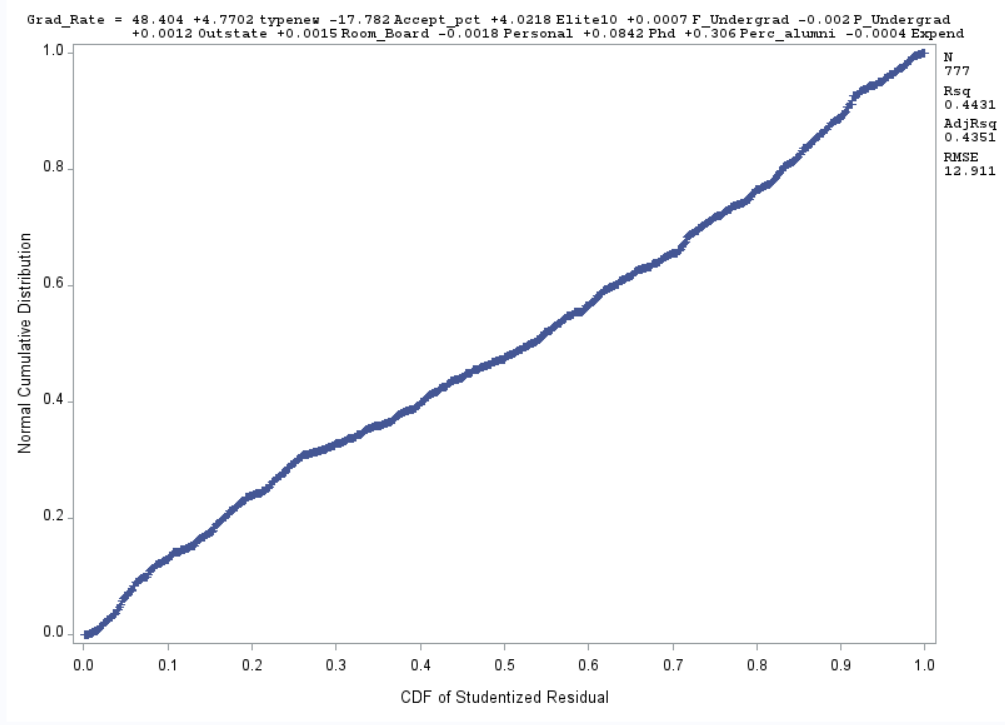
**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend;

plot npp.\*student.;

**run**;

The Plot is as follows:



We can make the following observation from the above plot:

* The slight curve at the top and bottom give evidence for the presence of outliers in the dataset.
* The line is mostly linear which means that the error is normally distributed over all the observations/predictions.

1. To study the presence of outliers we will make use of Residual values and for influential points we will make use of Cook’s D Value.

The code to generate the residual and Cook’s D value is as follows:

title "Residual and Influential point analysis";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend / influence r;

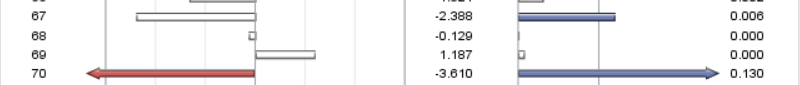
**run**;

The output which is generated is as follows:















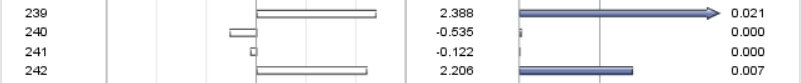






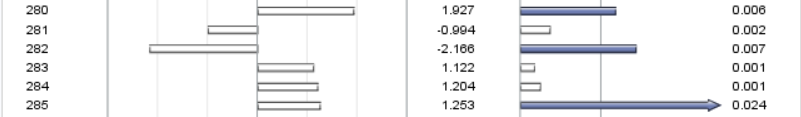




















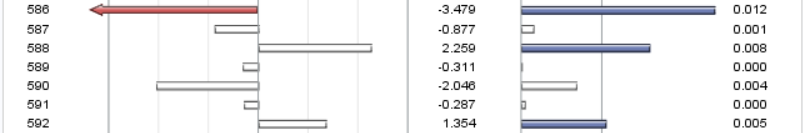






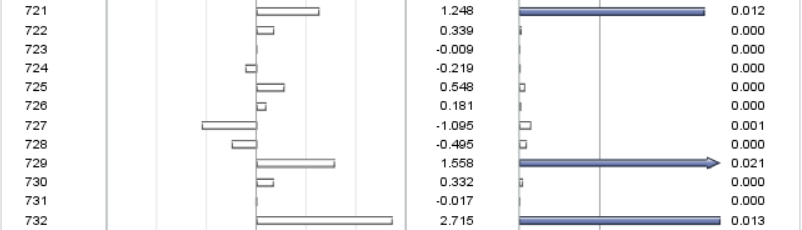
















As we can see the data points which have **Cook’s D Value greater than or equal to 0.005 will be considered as influential points**. There are many influential points present in this dataset. The influential points have a solid Blue Arrow/rectangle in the right half of the table.

The data points which have **Studentised Residual value greater than or equal to 3 will be flagged as outliers**. There are **8 outliers** present in this dataset which are represented by solid red color arrow/rectangle.

The points which have residual value greater than 3 and Cook’s D Value greater than 0.005 can be removed as they are possible outliers and are distorting the final equation but this is not always advisable.

The points whose Residual values are close to 3 and whose Cook’s value is just over 0.005 must be investigated.

1. The **Adjusted R-Squared value of this model is .4351** i.e. **43.51% of the variance** in the graduation rate can be explained using this model which has **11 predictors.**

This is a good model because the P-value associated with the **F-Statistic is also <0.0001** and hence we can say this is a good model but in practice I wouldn’t suggest this model because it explains very less variance in the Graduation Rate and hence it’s up to the business to decide on the model.

This model also doesn’t contain any insignificant variables.

1. Based on the Regression analysis that we performed **OutState is the strongest contributor** to the Graduation Rate, it alone can explain **~28.5% variance in the Graduation rate**.

As shown earlier in the boxplots, there is a significant difference in the graduation rate which is observed in the Private and Public universities as well as the Elite and non-Elite class universities.

**The mean graduation rate is higher in Private universities when compared to public universities.**

**The mean graduation rate is higher in Elite universities when compared to non-elite universities.**

To compute which is the strongest contributor we run the following piece of code:

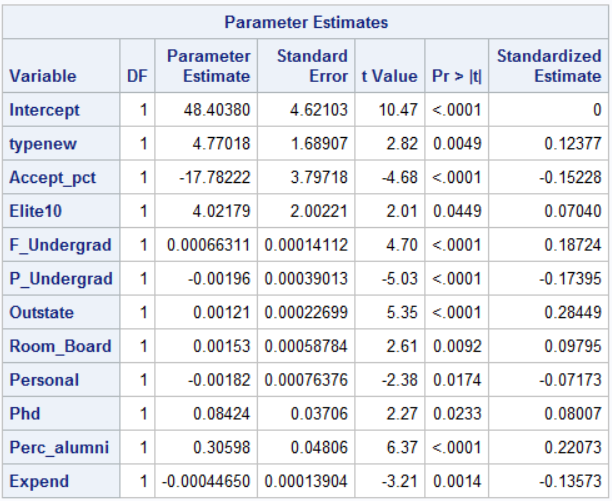
title "Residual and Influential point analysis";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend / influence r;

**run**;

This code generates the following output:



And we can see that OutState has the highest value in the Standardized Estimate value and hence it is the strongest contributor in our final model.

The Whole code for this dataset is as follows:

title "College Dataset with Dummy Variable for Private";

**data** college;

infile "college.csv" delimiter =',' missover firstobs=**2**;

input School $ Private $ Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend Grad\_Rate;

typenew = **0**;

if private = 'Yes' then typenew = **1**;

**proc** **print**;

**run**;

title "Univariate analysis on GradRate";

**proc** **univariate** normal;

var Grad\_Rate;

histogram / normal (mu = est sigma = est);

**run**;

title "Scatterplot Matrix";

**proc** **sgscatter**;

MATRIX Grad\_Rate typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend;

**run**;

title "Correlation values";

**proc** **corr**;

var Grad\_Rate typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend;

**run**;

title "Individual scatter plots";

**proc** **plot**;

plot Grad\_Rate\*(typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend);

**run**;

title "BoxPlot of GradRate vs Public/Private ";

**proc** **sort**;

by Private;

**run**;

**proc** **boxplot**;

plot Grad\_Rate\*Private;

**run**;

title "BoxPlot of GradRate vs Elite";

**proc** **sort**;

by Elite10;

**run**;

**proc** **boxplot**;

plot Grad\_Rate\*Elite10;

**run**;

title "Full Model";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend / vif stb;

**run**;

title "New model using Forward selection method";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend / selection = forward sle = **0.05** sls = **0.05**;

**run**;

title "New model uaing adjusted R2 selection method";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal Phd Terminal S\_F\_Ratio Perc\_alumni Expend / selection = adjrsq sle = **0.05** sls = **0.05**;

**run**;

title "Final Model";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend / vif stb;

**run**;

title "Studentised values vs Predicted values Plot";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend;

plot student.\*predicted.;

**run**;

title "Normality Plot";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend;

plot npp.\*student.;

**run**;

title "Residual and Influential point analysis";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend / influence r;

**run**;

title "Strongest contributor";

**proc** **reg**;

model Grad\_Rate = typenew Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal Phd Perc\_alumni Expend / stb;

**run**;