**Aamir Patel**

**Data Analysis And Regression**

**Assignment-5** | **Total Points: 25**

Note:

* All assignments should be submitted in a **single MS WORD format**, no PDFs or any other file types will be accepted. If you submit any other file type, it will not be graded.
* No extensions will be given unless for a documented reason specified in the syllabus, no late assignments past the due date even a couple of minutes late will be accepted as you have an extra day (8-days) to submit your assignments.
* Submitting work that is not yours is grounds for an automatic ‘F’ for the entire course – this includes taking content and ideas from others or consulting others to complete your deliverables other than your instructor.
* SAS software and virtual server stalls, gets slow and crashes; so start early and keep multiple backups in multiple places/mediums. Late submission or inability to do the assignment due to server and/or software issues will not be accepted. Any issues relating with SAS, contact IS using the phone number provided in the syllabus, I won’t be able to help you with DePaul software related issues.

***Note: For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.***

**Problem 1 [5 pts]** You will continue the prediction, confidence interval and prediction interval for the **banking** dataset that was analyzed in Assignment 4. Since you would have altered the dataset to exclude outliers/influential points and/or multicollinearity, use the dataset and the code that was used to generate your final model. Note: Make sure you rerun the whole banking code from assignment 4, before you do this last part.

1. Use the **fitted regression model from Assignment 4** to predict the average bank balance for a specific zip code area where there is a plan to open a new branch. Census data in that area show the following values: median age is 34 years, median education is 13 years, median income is $89,000, median home value is $160,000, median wealth is 140,000. Using SAS, compute the predicted average bank balance, 95% confidence interval and prediction interval for your estimate. Make sure to use SAS coding to determine the values. Include all relevant outputs. Discuss your findings.

* The Regression model from Assignment 4 is: -

Balance = -11867 + 340 (Age) + 770 (Education) + 0.02485 (HomeVal) + 0.11021 (Wealth)

So if we want to predict Balance of someone with the median Age of 34, Education of 13, Income of $89,000, HomeVal of $160,000, and Wealth of $140,000.

For Age = 340 \* 34 = $11,560

Education = 770 \* 13 = $10,010

Income = 89,000 (Not significant)

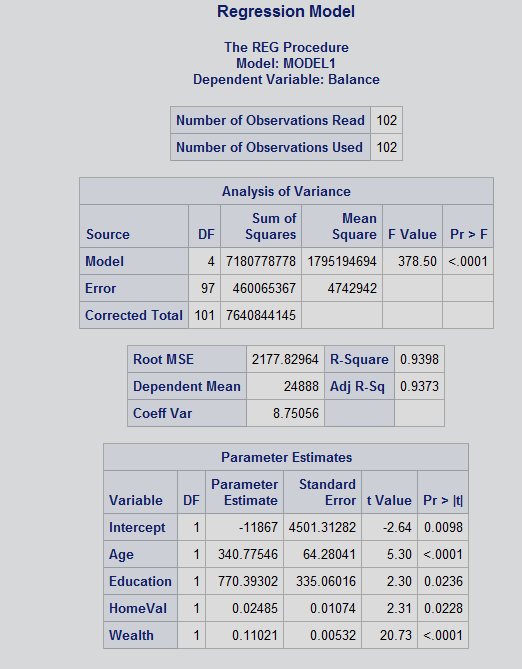
HomeVal = 0.02485 \* 160,000 = $3,976

Wealth = 0.011021 \* 140,000 = $1,542.94

Therefore Predicted Balance would be

Balance = -11867 + $11,560 + $10,010 + $3,976 + $1,542.94

= $15,221.94. with a confidence interval of 95%.



Copy and paste your FULL SAS code into the word document along with your answers.

**DATA** BankingFull;

**PROC** **IMPORT** datafile="Bankingfull.txt"

out=Bankingfull replace;

delimiter='09'x;

getnames=yes;

datarow=**2**;

**run**;

Title "Regression Model";

**PROC** **REG**;

model Balance = Age Education HomeVal Wealth / influence r;

plot student.\*(Age Education HomeVal Wealth predicted.);

plot npp.\*student.;

**run**;

TITLE " New Regression Model with requirements";

**DATA** NEW;

INPUT Age Education HomeVal Wealth;

DATALINES;

**34** **13** **160000** **140000**;

**proc** **print**;

**run**;

title"Regression Model for Full Model";

**PROC** **REG** CORR;

MODEL Balance = Age Education Income HomeVal Wealth/stb influence r;

title "Reduced Mode";

MODEL Balance =Age Education Income Wealth / stb influence r;

title "Residual Analysis";

plot residual.\*(predicted.Age Education Income Wealth

plot npp.\*residual;

**RUN**;

**PROBLEM 2 [20 pts]**

This problem asks you to build a model for the college dataset (college.csv) that contains the following variables:

*School School name*

*Private public/private indicator. YES if university is private, NO if university is public.*

*Accept.pct percentage of applicants accepted*

*Elite10 Elite schools with majority of students from the top 10% of their high school class*

*(0- Not Elite, 1-Elite)*

*F.Undergrad number of full-time undergraduate students*

*P.Undergrad number of part-time undergraduate students*

*Outstate Out-of-state tuition*

*Room.Board room and board costs*

*Books estimated book costs*

*Personal Estimated personal spending*

*PhD Percent of faculty with PhD*

*Terminal Faculty with terminal degrees (terminal degree is a university degree that is either*

*highest on the academic track or highest on the professional track in a given field*

*of study)*

*S.F.Ratio Student/faculty ratio*

*perc.alumni Percent of alumni who donate*

*Expend Instructional expenditure per student*

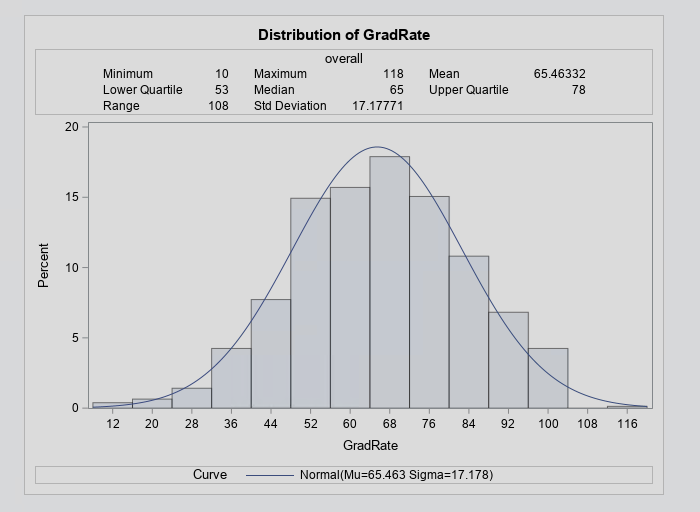
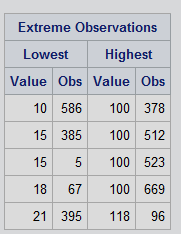
*Grad.Rate Graduation rate in 4 years*

Apply regression analysis techniques to analyze the relationship among the observed variables and build a model to predict Graduation Rates (Grad.Rate). **Note: Depending on how you import you data (INFILE or IMPORT) the SAS may relabel the column names. Make sure to use the variable names that appear when you use a proc print.**

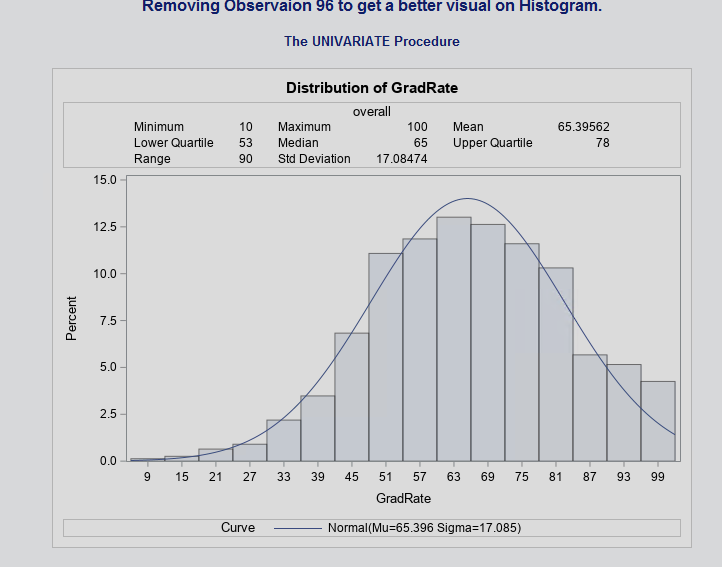
***Note: Before you start, open the college.csv file, and examine the data.***

Answer the following questions.

1. Analyze the distribution of Grad.Rate and discuss if the distribution is symmetric, or if you need to apply any transformation (This is the data exploration stage, therefore use the appropriate statics to explore your data).

* Initially the distribution looked normal, but I noticed that there was 1 variable that was making it seem like it wasn’t, so after confirming that is acting as an influential point, I removed the variable 96.
* After removing the variable 96, I can confirm that there is a bit of skewness to the right. So, this dataset needs to be transformed for us to continue.
* 
* 

Results after Removing variable : -

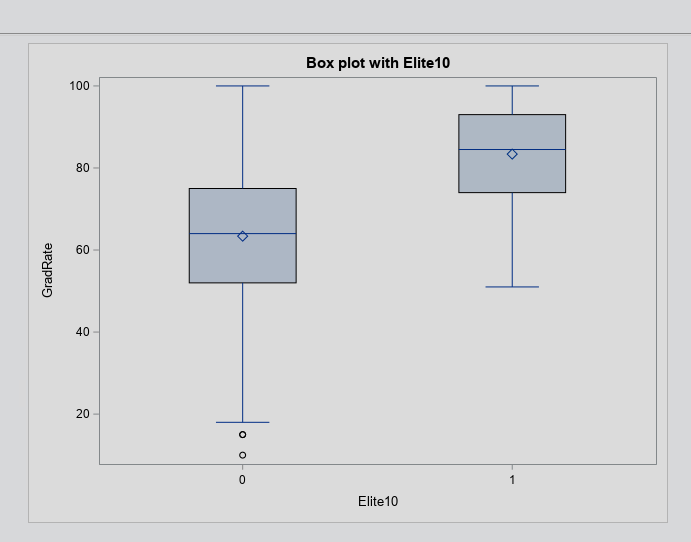
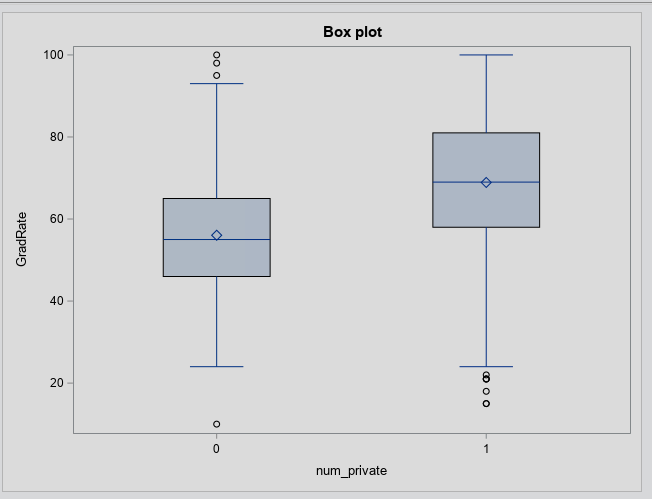


1. Create scatterplots for Grad. Rate vs each of the independent variables. What conclusions can you draw about the relationships between Grad. Rate and the independent variables? (No need to include the scatterplots in your submission).

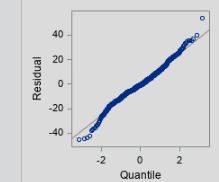
* Upon analyzing scatterplots, I can confidently say this for each of the independent variables.
* Grad Rate and Acceptance Rate shows a weak nonlinear negative association at -0.292.
* Grad Rate and Elite 10 shows a weak positive nonlinear association at +0.352.
* Grad Rate and F.Undergrad shows a weak negative nonlinear association at -0.077.
* Grad Rate and P.Undergrad shows a weak negative nonlinear association at 0.256.
* Grad Rate and Outstate shows a neutral positive nonlinear association at 0.575.
* Grad Rate and RoomBoard shows a neutral positive nonlinear association at 0.425
* Grad Rate and Books shows a weak negative nonlinear association at -0.00015
* Grad Rate and Personal has a weak negative nonlinear association at -0.266
* Grad Rate and PHD has a weak positive nonlinear association at 0.321
* Grad Rate and Terminal has a weak positive nonlinear association at 0.301
* Grad Rate and SFRatio shows a weak negative nonlinear association at -0.308
* Grad Rate and PercAlumni has a neutral positive nonlinear association at 0.494
* Grad Rate and Num\_private has a weak positive association at 0.335.

1. Build boxplots to evaluate if graduation rates vary by university type (private vs public) and by status (elite vs not elite). Include the boxplots and discuss your findings. (See SAS Procedures section on D2L if you need the code to generate a boxplot).

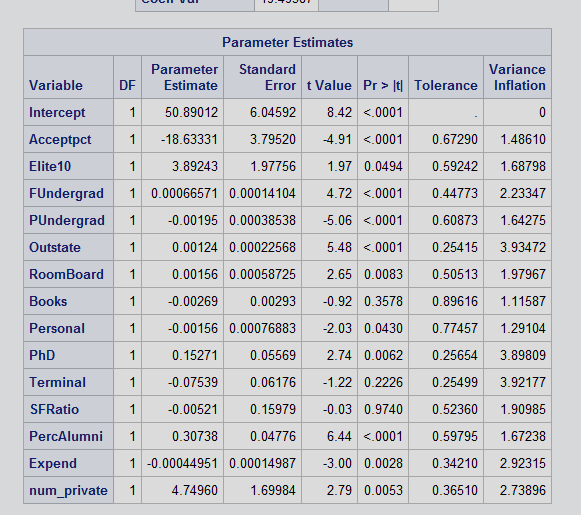
* Upon analyzing the box plots for each num\_private and Elite10 for GradRates, we can see that graduation rate for the students who attend private university is greater than non private universities. We can see the same results in Elite10, Students who are part of Elite schools have a higher graduation rate than students who don’t attend Elite schools.



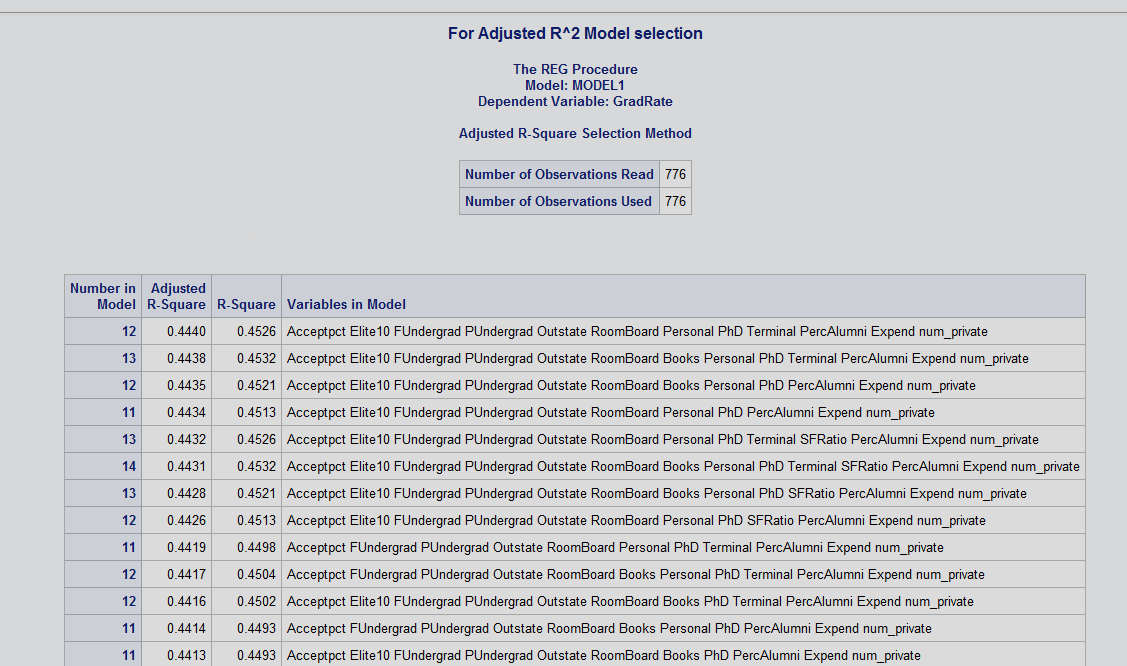
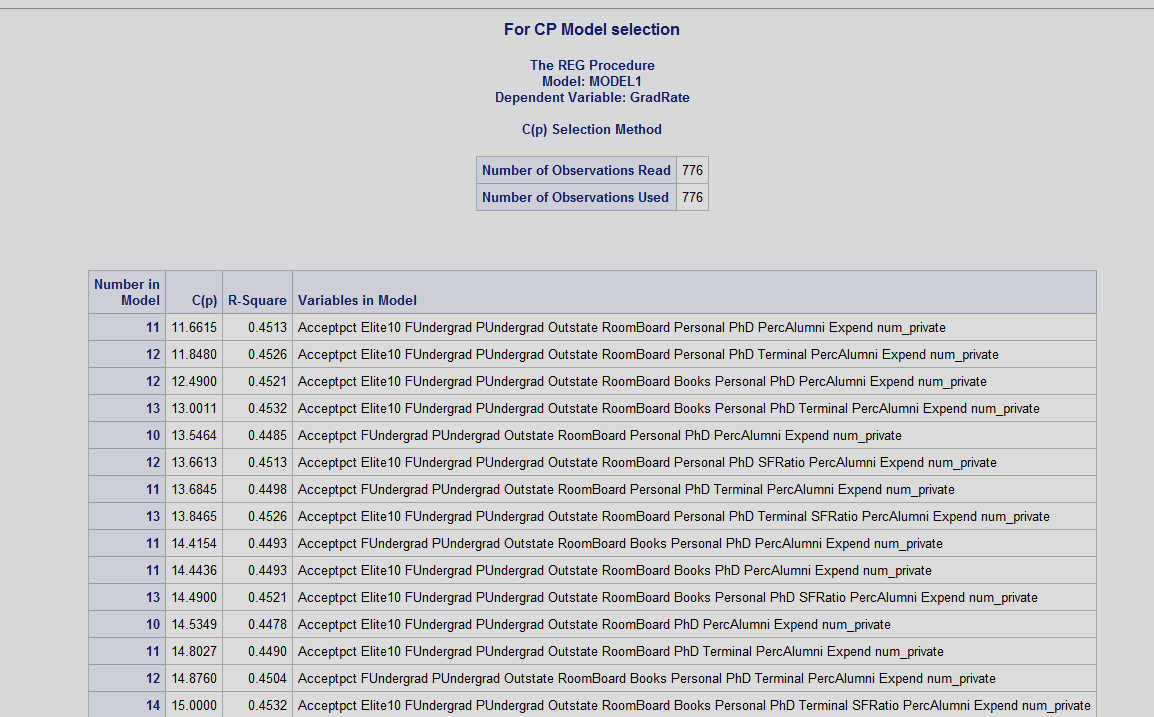
1. Fit a full model (with all independent variables) to predict Grad.Rate. Discuss the parameter estimates, significance, goodness-of-fit and AdjR2 values. Include the relevant output.

* Looking at the model, we can see that the adjusted R^2 rate is rather low, at just 44.3%. and parameter estimates are all rather low as well. For the goodness of fit test,
* H0 = There is no relationship,
* Ha = There is a relationship.
* In this case, we can confirm that there is a relation ship and rule of h0, since residual plot has a almost perfect slope to it.
* 
* 

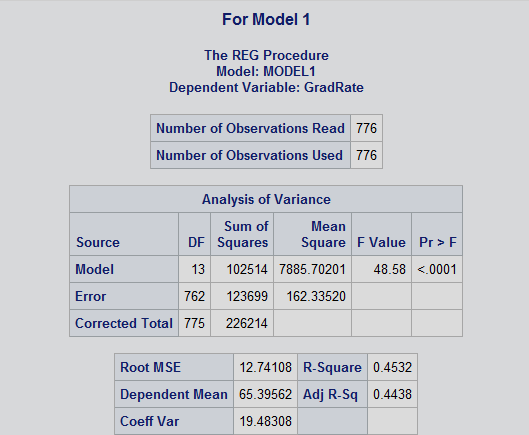
1. Does multi-collinearity seem to be a problem here? What is your evidence? Compute and analyze the VIF statistics. Include the relevant output and discuss your answer.

* There does not seem to be any issues with multi-collinearity, which is confirmed by VIF and Tolerence test in SAS. Our findings showed that there were no variables with a VIF greater than 10, or tolerance less than 0.10.
* 

1. Apply TWO variable selection procedures to find an optimal subset of independent variables to predict Grad.Rate*.* You can choose any two procedures among the ones we learned in class: backward selection, forward selection, adj-R2, Cp, stepwise. Make sure to include the o/p of the 2 selection methods. No need to discuss the models, include the outputs.

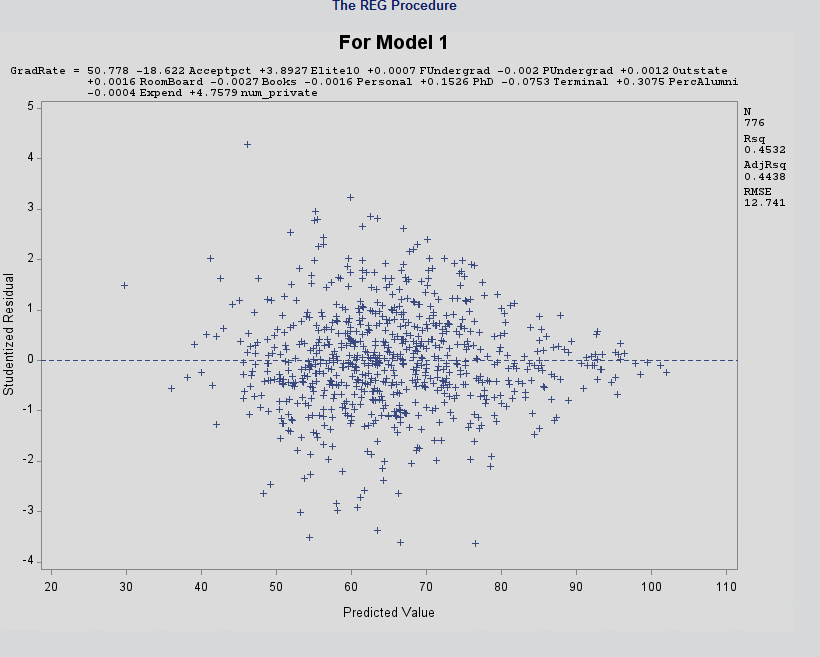
* For my variable selection, I selected adj-R2 and Cp.
* Adjusted R^2
* 
* CP
* 

1. Fit a final regression model **M1** for Grad.Rate based on the results in f) – i.e. optimal model. Explain your choice. Write down the expression of the estimated model **M1**.

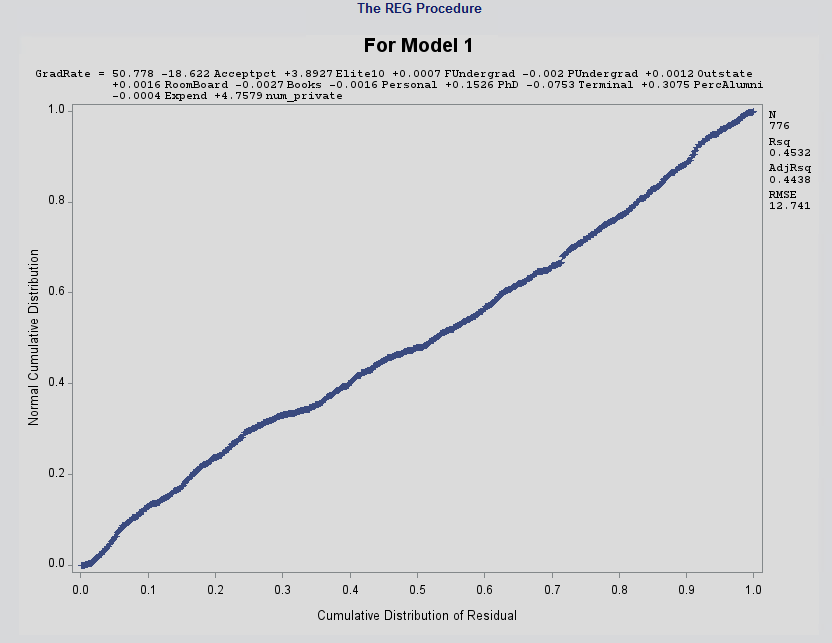
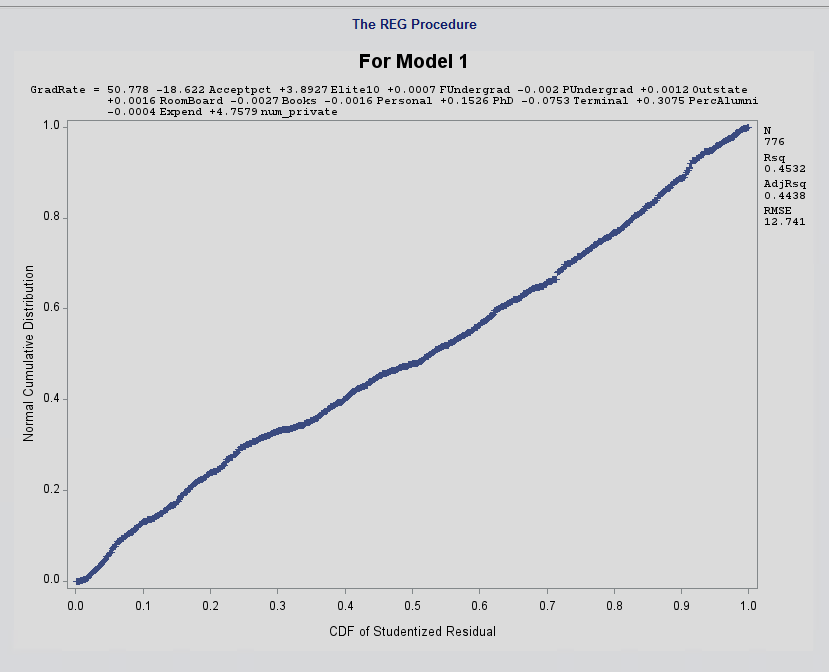
* I selected the model with the most variables, and the highest R ^2, so I could get a better estimated result. And the expression for the estimated model is.
* **Grad Rate = 50.77 + (-18.62 Accept pct) + 3.89 (Elite 10) + 0.00066 (F Undergrad) + -0.00195 (P Undergrad) + 0.0012 (outstate) + 0.00156 (Room Board) + (-0.00269 Books) + (-0.00156 Personal) + 0.15261 (PhD) + (-0.0753 Terminal) + 0.30749 (Perc Alumni) + (-0.000447 Expend) + 4.75793 (num\_private).**
* 
* 

1. Draw a plot of the studentized residuals against the predicted values. Does the plot show any striking pattern indicating problems in the regression analysis? Include the outputs and explain.

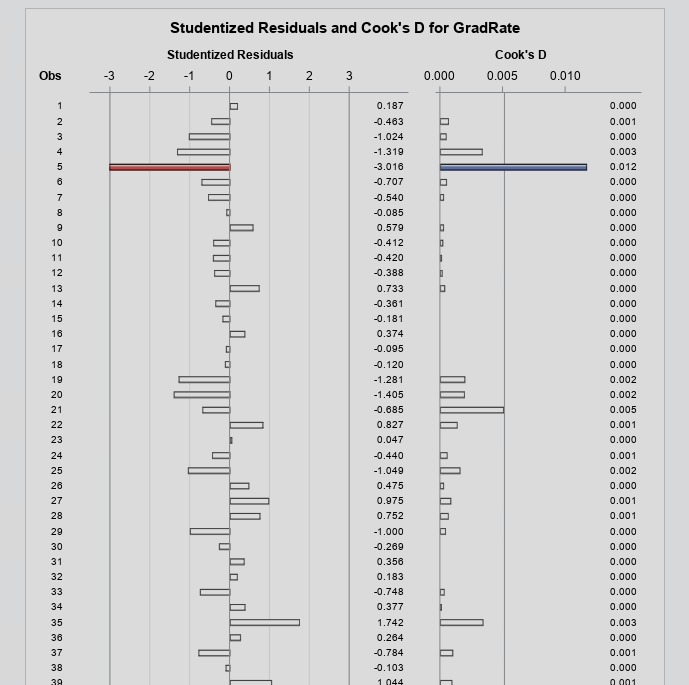
* Yes the plot seems to be showing a fish like pattern, which is giving me a sign that something is a little wrong.



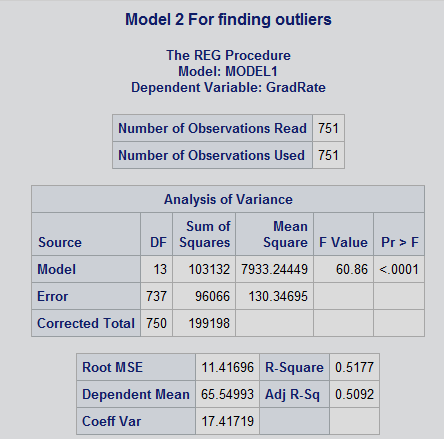
1. Analyze normal probability plot of residuals. Is there any evidence that the assumption of normality is not satisfied? Include the outputs and explain.

* No, everything in the NPP plot seems normal, and no obvious issues can be seen.
* 
* 

1. Are there any outliers or Influential Points? Compute appropriate statistics. Include the outputs. Take any action you think is necessary and explain why/why not you took these actions?

* Yes there are many outliers and influential points, I will be removing them so that our observations are more accurate, and not being influenced by some mistakes.
* 

1. Analyze the AdjR2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities.

* We can see that after removing those influential points and outliers, our adjusted R^2 jumped up over 6% at 50.9%, indicating they were pulling back out adjusted R^2. It also shows that there is a strong enough relationship between Graduate Rate and other variables.
* 

1. Draw conclusions on graduation rates based on your regression analysis. What are the most important predictors in your model? Does your model show a significant difference in graduation rates between private and public universities? Do “elite” universities have higher graduation rates? Explain.

* After doing the analysis I can confidently say that variables that has the most impact on graduation rate are as follows Elite10, Outstate, FUndergrad, PercAlumni, and PhD. Others does not have a major a impact on the Graduation Rate. Also, Private and Public Universities does have a significant impact on Graduation Rate. No elite10, has a minor effect on graduation rate.

1. Copy and paste your FULL SAS code into the word document along with your answers.

\*Importing Dataset;

**DATA** college;

Infile "College.csv" firstobs = **2**

delimiter = ',';

input school$ Private$ Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal SFRatio PercAlumni Expend GradRate;

\*Setting Local Variables;

if(Private = 'Yes') then num\_private = **1**;

else num\_private = **0**;

**RUN**;

\*Printing out the DATASET;

**PROC** **PRINT** data = college;

**run**;

\*Printing Histogram;

title "Printing Histogram without any changes";

**PROC** **UNIVARIATE** normal ;

var GradRate;

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

INSET min max mean Q1 Q2 Q3 Range stddev/ header = 'overall'

pos=tm;

**run**;

\*Removing Observaion 96 to get a better visual on Histogram.;

title "Removing Observaion 96 to get a better visual on Histogram.";

**data** college;

set college;

if \_n\_ = **96** then delete;

**run**;

**PROC** **UNIVARIATE** normal ;

var GradRate;

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

INSET min max mean Q1 Q2 Q3 Range stddev/ header = 'overall'

pos=tm;

**run**;

\*Creating scatter plots;

**proc** **sgscatter**;

title"Scatterplots 1";

matrix GradRate Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books ;

**run**;

**proc** **sgscatter**;

title"Scatterplots 2";

matrix GradRate Personal PhD Terminal SFRatio PercAlumni Expend num\_private;

**run**;

\*proc reg;

**proc** **reg**;

title"Regression model";

model GradRate = Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal SFRatio PercAlumni Expend num\_private;

**run**;

\*Correlation;

**proc** **corr**;

title "Correlation";

var GradRate Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal SFRatio PercAlumni Expend num\_private;

**run**;

\*Box plots;

**proc** **sgplot** data= college;

title "Box plot";

vbox GradRate / category = num\_private;

**run**;

\*Box plots wioth elite;

**proc** **sgplot** data= college;

title "Box plot with Elite10";

vbox GradRate / category = Elite10;

**run**;

\*For Multicollinearity;

**proc** **reg**;

title "Multicolinearity";

model GradRate = Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal SFRatio PercAlumni Expend num\_private / vif tol;

**run**;

\*For Adjusted R^2 Model selection;

**proc** **reg**;

title "For Adjusted R^2 Model selection";

model GradRate = Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal SFRatio PercAlumni Expend num\_private / selection=adjrsq;

**run**;

\*For CP Model selection;

**proc** **reg**;

title "For CP Model selection";

model GradRate = Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal SFRatio PercAlumni Expend num\_private / selection=CP;

**run**;

\*For Selected selection;

**proc** **reg**;

title "For Model 1";

model GradRate = Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal PercAlumni Expend num\_private;

plot student.\*predicted.;

plot npp.\*residual.;

plot npp.\*student.;

**run**;

\*For Finding outliers;

**proc** **reg**;

title "Model 1 For finding outliers";

model GradRate = Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal PercAlumni Expend num\_private / influence r;

**run**;

\*Removing outliers and Influential points;

**data** college;

set college;

if \_n\_ in (**70** ,**64**, **5**, **47**, **67**, **98**, **113**, **126**, **134**, **142**, **152**, **169**, **197**, **201**, **238**, **215**, **264**, **284**, **317**, **319**, **377**, **394**, **506**, **640**, **728**) then delete;

**run**;

\*For Finding outliers;

**proc** **reg**;

title "Model 2 For finding outliers";

model GradRate = Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal PercAlumni Expend num\_private / stb;

**run**;