**Homework 3 - Jon Doretti**

**DATA ANALYSIS AND REGRESSION**

**Assignment-3** | **Total Points: 25 pts for DSC 323; 30 pts for DSC 423**

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 (7-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.

• **Make sure to double check your submissions. After you submit the assignment, log out of D2L, log back in, and click on your submission to see if you submitted the right file(s) and it is the correct version. Wrong submissions will not be graded.**

***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 [25 pts] – to be answered by everyone**

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*

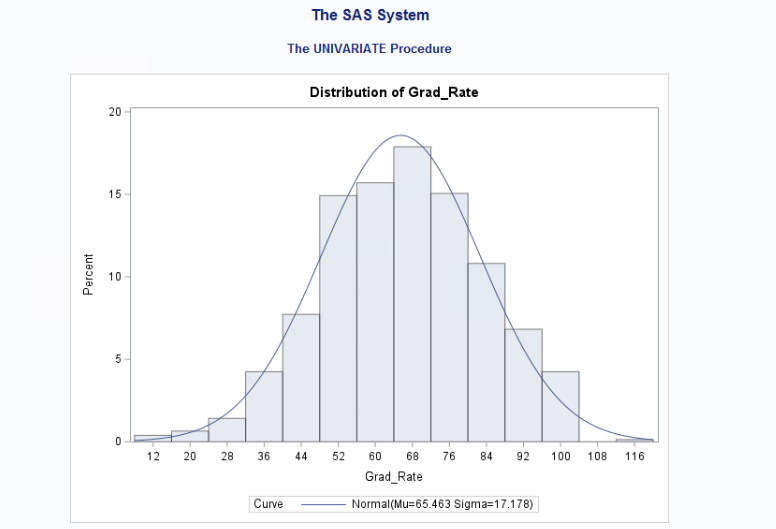
*Grad.Rate Graduation rate in 4 years*

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.

a) 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). 

The histogram above shows that there is an error in the data set. No university can have a graduation rate of above 116%. This would mean that more students are graduating than exist at the university. With the error - the histogram has a normal distribution. However by analyzing the graph, with the removal of the error, the histogram will either become slightly left skewed or moderately left skewed.

b) 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).

Association = Grad Rate -> x, where x = any variable

Grad rate -> Accept\_Pct: low negative linear

-> F\_Undergrad: low negative linear

-> P\_Undergrad: low negative linear

-> Outstate: normal positive linear

-> Room\_Board: normal positive linear

-> Outstate: normal positive linear

-> Books: low positive linear

-> Personal: low negative linear

-> PhD: normal positive linear

-> Terminal: normal positive linear

-> S\_F\_Ratio: low negative linear

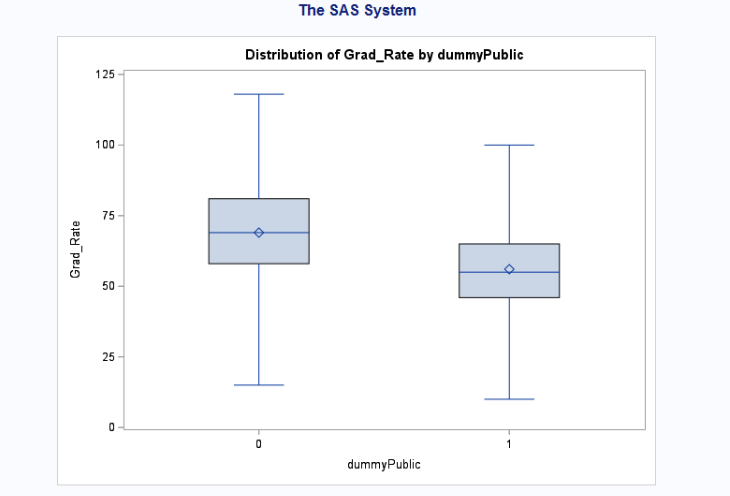
-> perc\_alumni: normal positive linear

-> Expend: normal positive linear

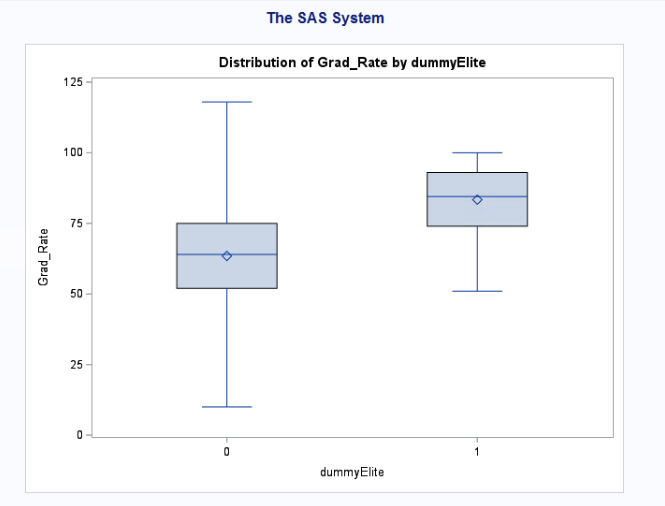
-> Private: Two vertical linear lines; 1or 0

-> Elite10: Two vertical linear lines; 1 or 0

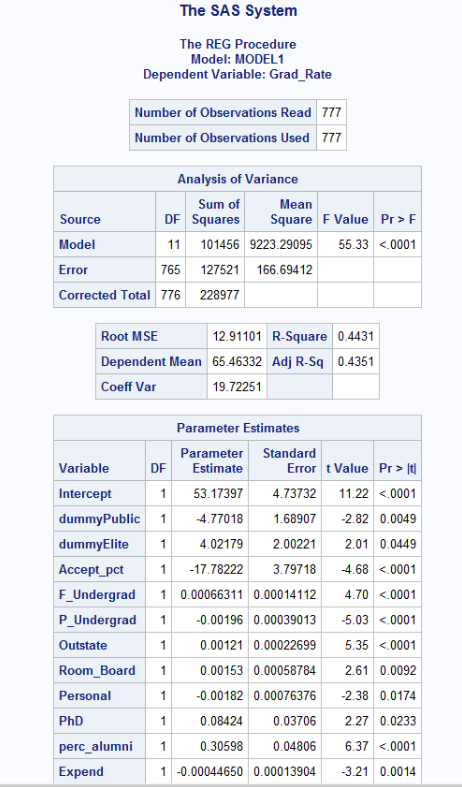
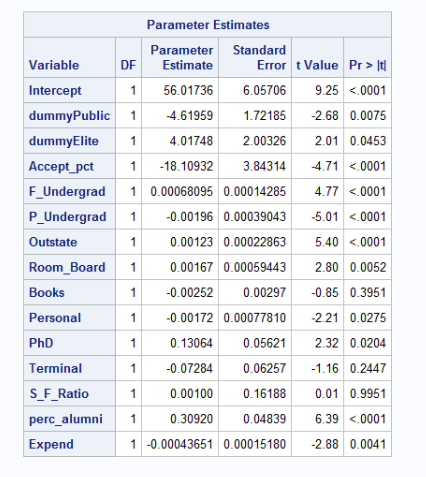
c) Build boxplots to evaluate if graduation rates vary by university type (private vs public) and by status (elite vs not elite). Include the boxplots and discuss your findings. (See SAS Procedures section on D2L if you need the code to generate a boxplot).



In this graph - you can tell that public schools (0) have a higher, all around, graduation rate than private schools (1). Examples for comparison would include Q1, the median and Q3.



In this graph - non-elite schools (0) have an overall lower graduation rate than elite schools (1). Examples for comparison would include Q1, the median and Q3.

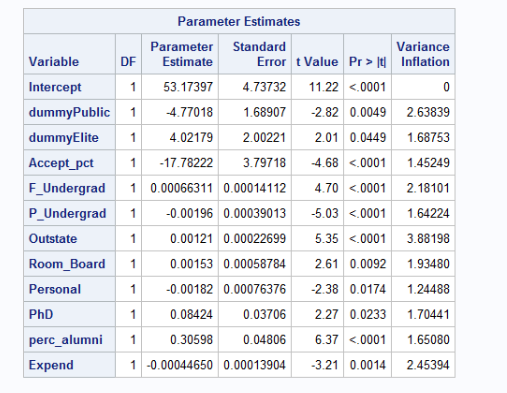
d) Fit a full model (with all independent variables) to predict Grad.Rate. Discuss the parameter estimates, significance, goodness-of-fit and AdjR2 values. Include the relevant output. 

The first graph includes all variables before removing non-significant variables based on the alpha test of .05. This test removes three variables: (1) S\_F\_Ratio, (2) Books, (3) Terminal.

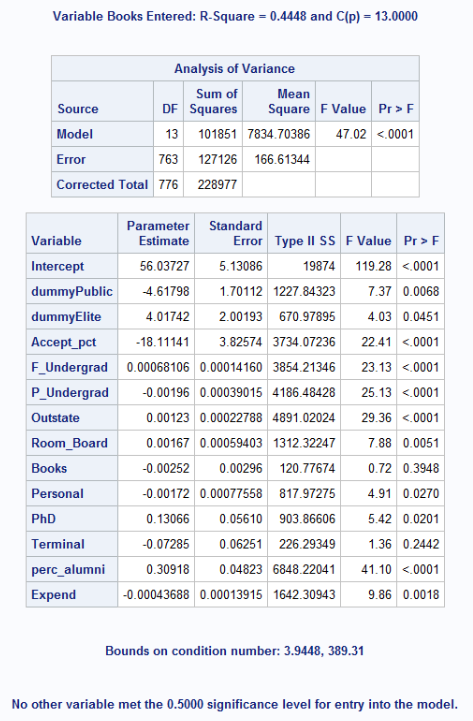
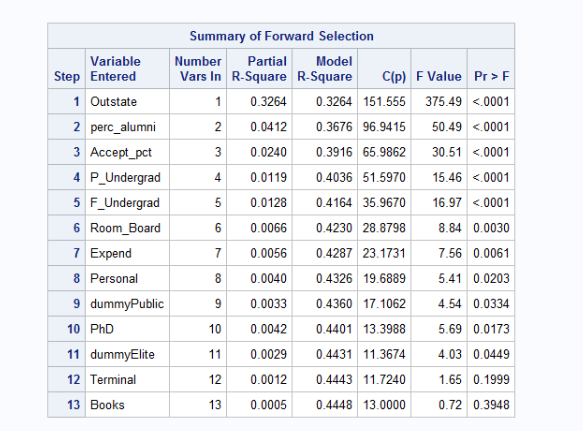
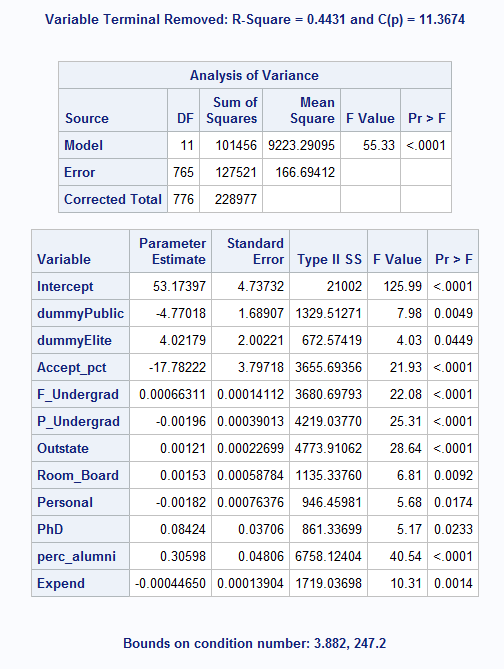
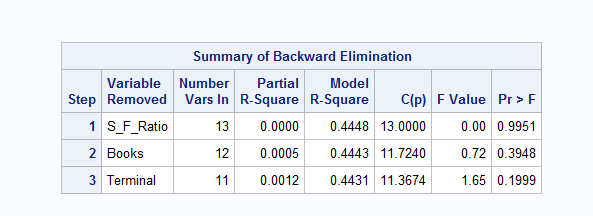
Full Model: Grad\_rate = 56.012 - 4.62(dummyPublic) + 4.02(dummyElite) -18.11(Accept\_pct) +.00068(F\_undergrad) - .002 (P\_Undergrad) + .0012(Outstate) + .0017(Room\_Board) - .0025(Books) - .0017(Personal) + .13(PhD) - .073(Terminal) + .001(S\_F\_Ratio) + .31(perc\_alumni) - .00044(Expend)

Based on the F Value (55.33) and the P Value (<.0001), Grad\_Rate and other variables have low to moderate association.

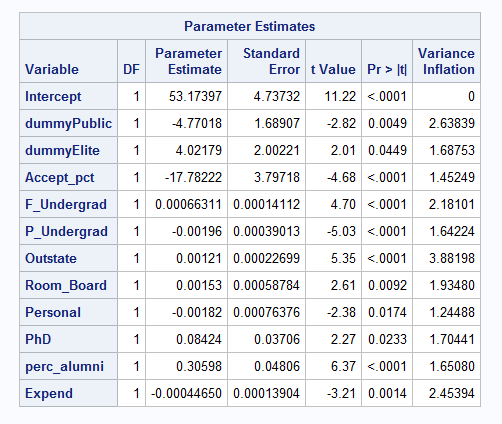
The Adjusted R Squared (.4352 or 43.51%) is not a strong value.

e) 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. 

Based on the above chart, there is no multicollinearity because the VIF, for all variables, is less than ten.

f) 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. 

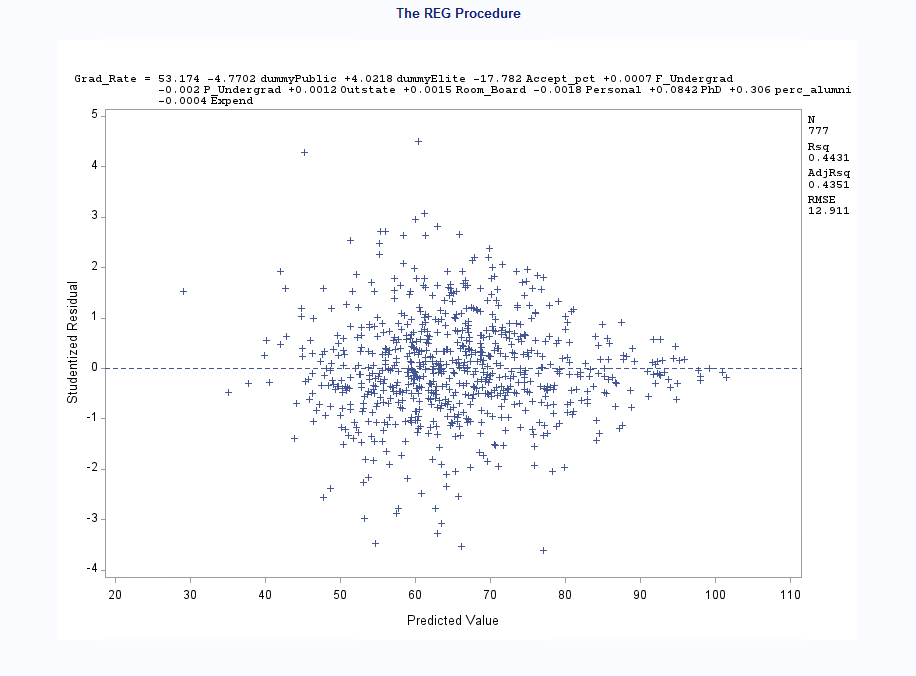
g) Fit a final regression model **M1** for Grad.Rate based on the results in f) – i.e. optimal model. Explain your choice. Write down the expression of the estimated model **M1**.



Similar to d) and e) - there is no multicollinearity because all VIF values are less than ten, for all variables.

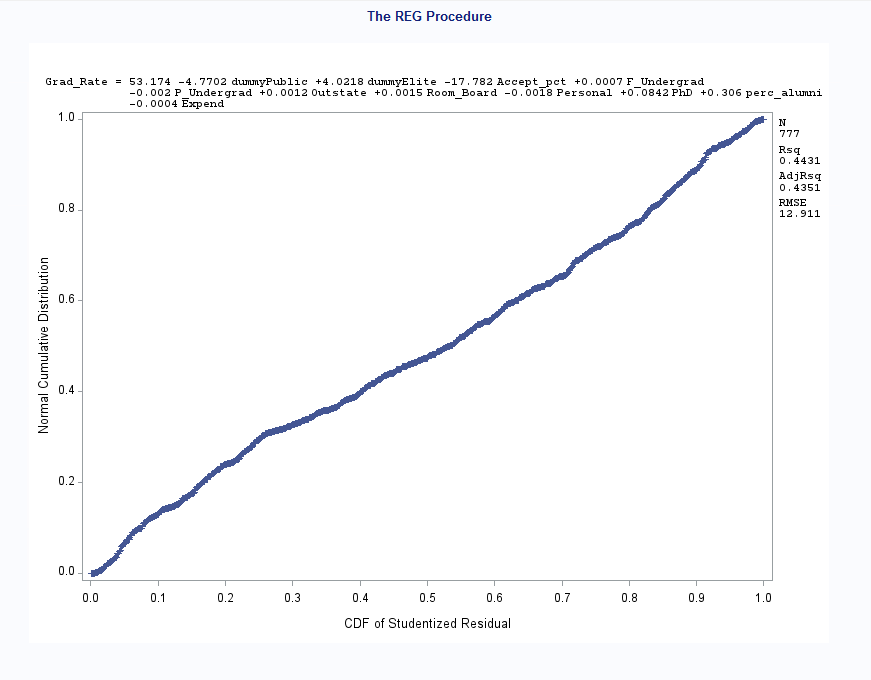
Full Model: Grad\_rate = 56.012 - 4.77(dummyPublic) + 4.02(dummyElite) -17.78(Accept\_pct) +.00066(F\_undergrad) - .002 (P\_Undergrad) + .0012(Outstate) + .0015(Room\_Board) - .0018(Personal) + .084(PhD) + .31(perc\_alumni) - .00045(Expend)

h) 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.



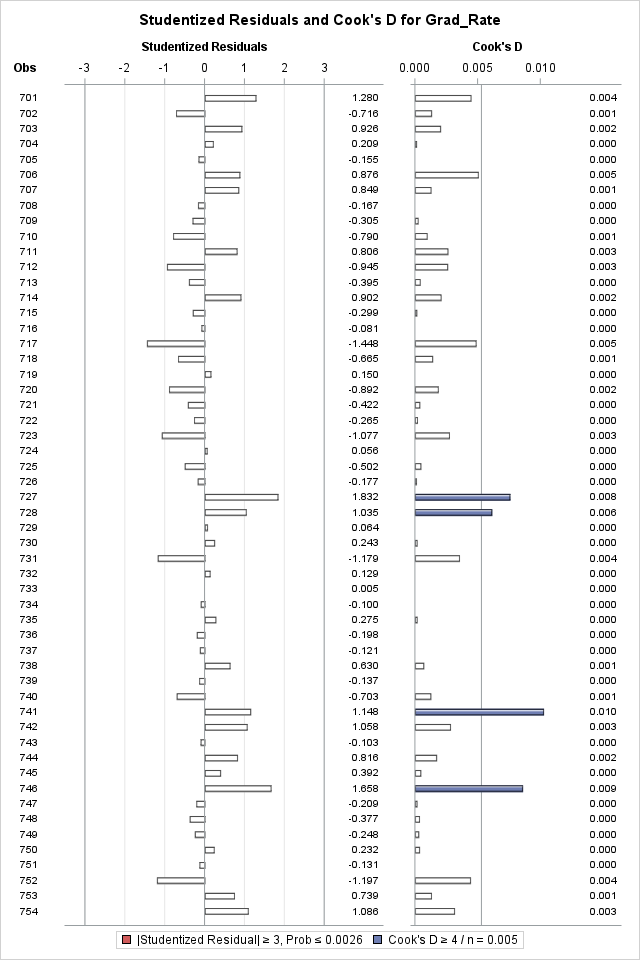
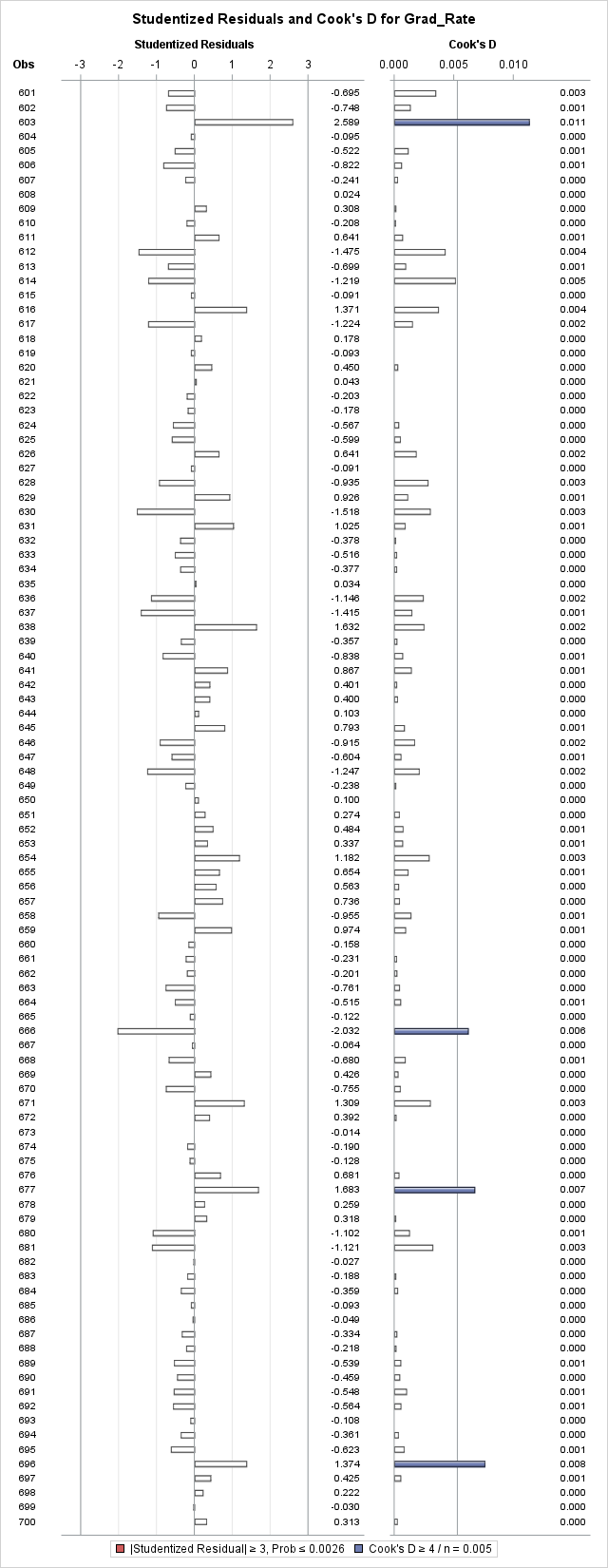
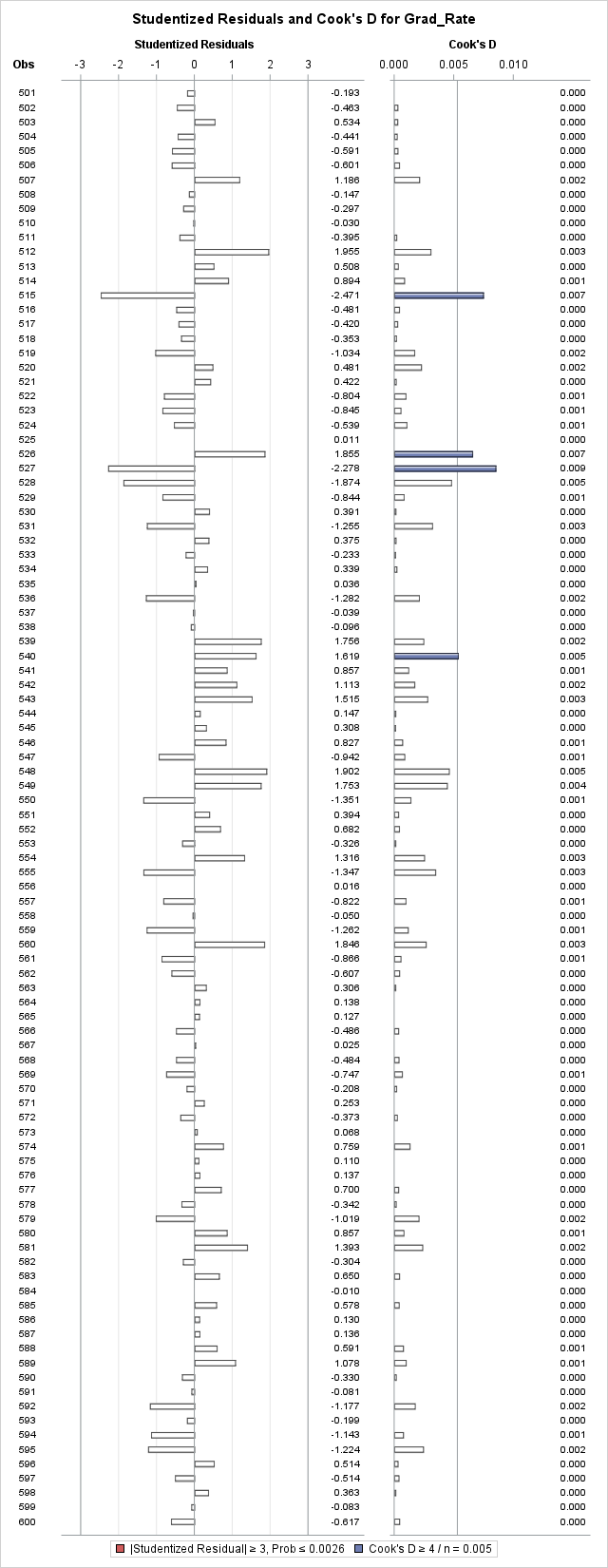
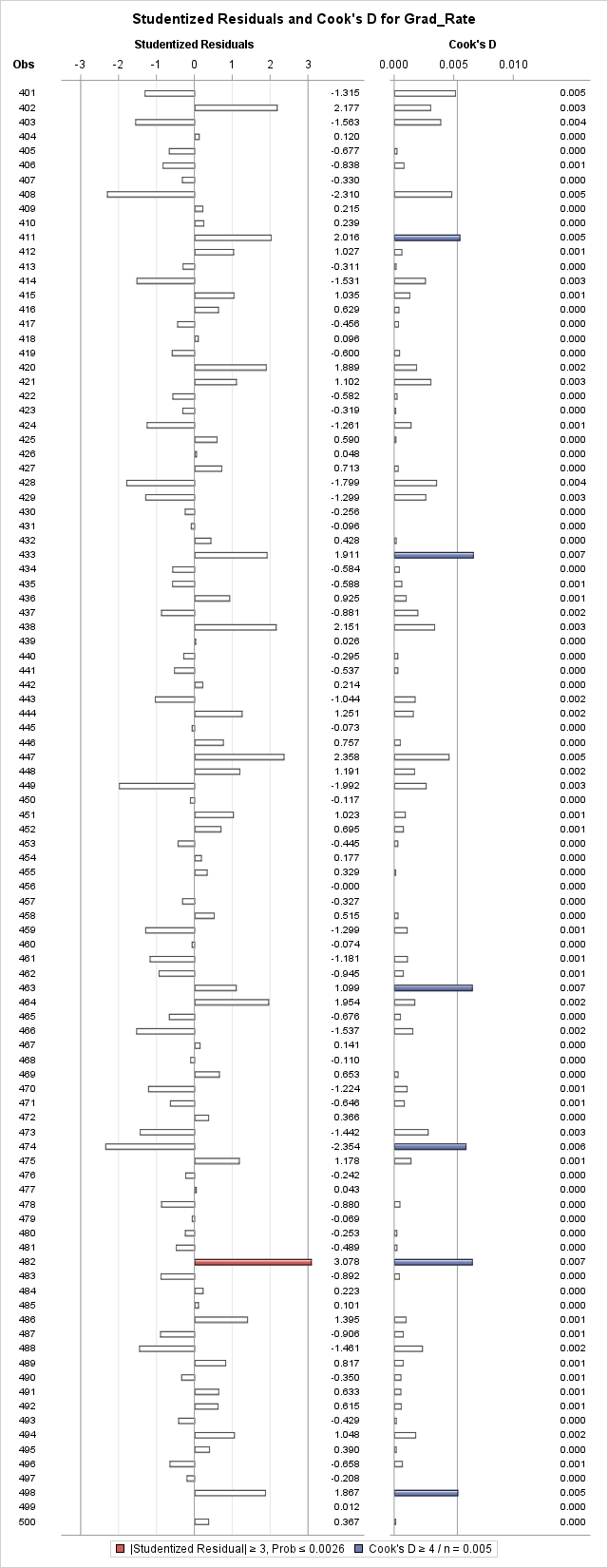
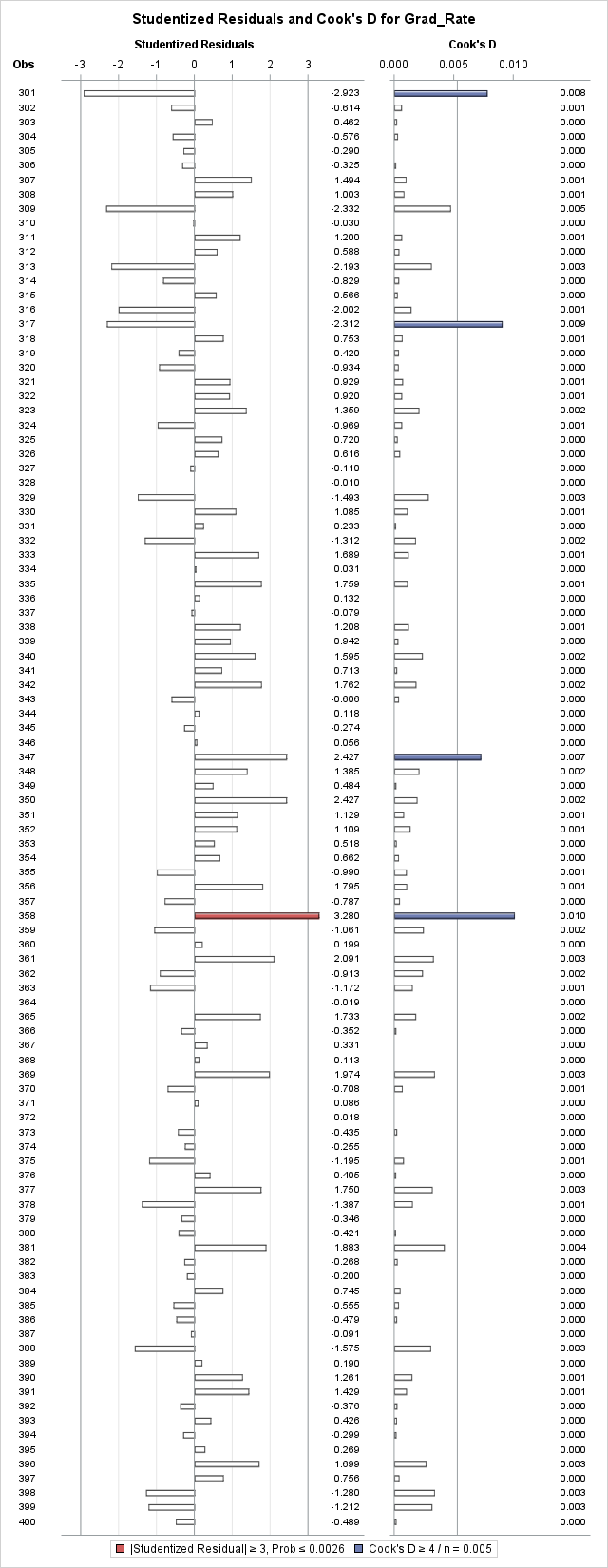
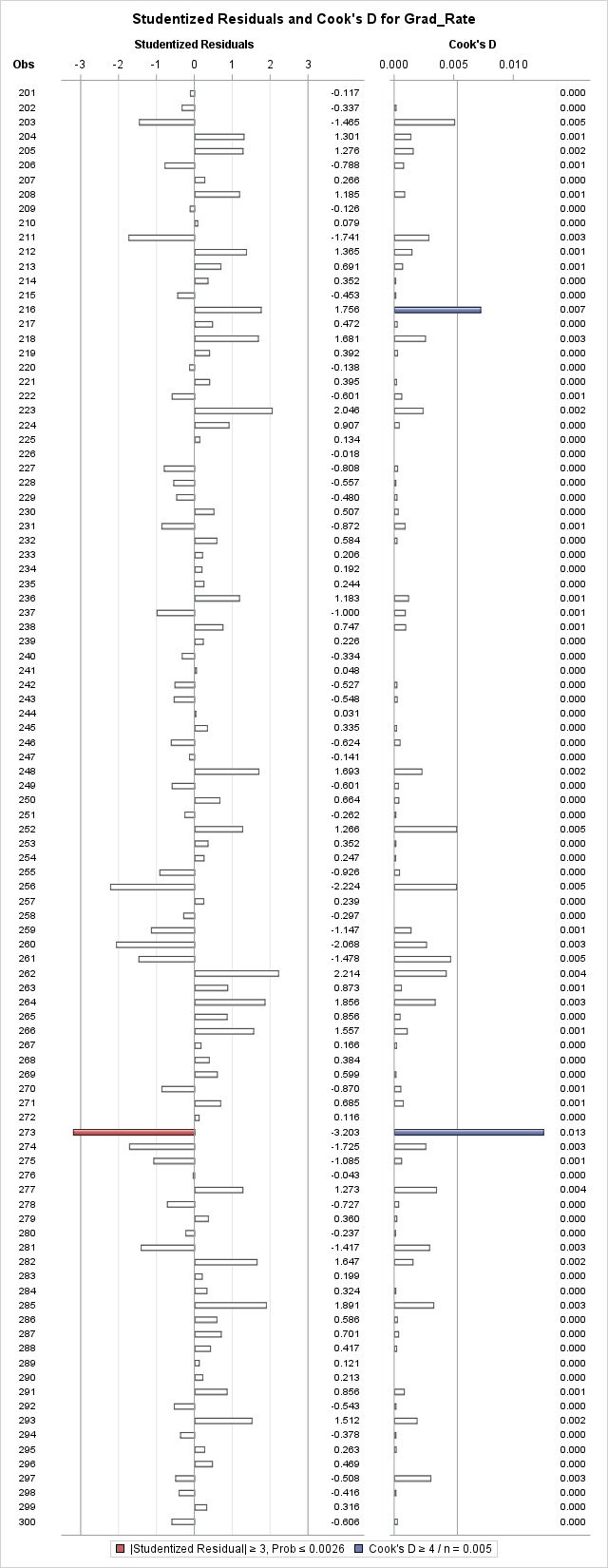
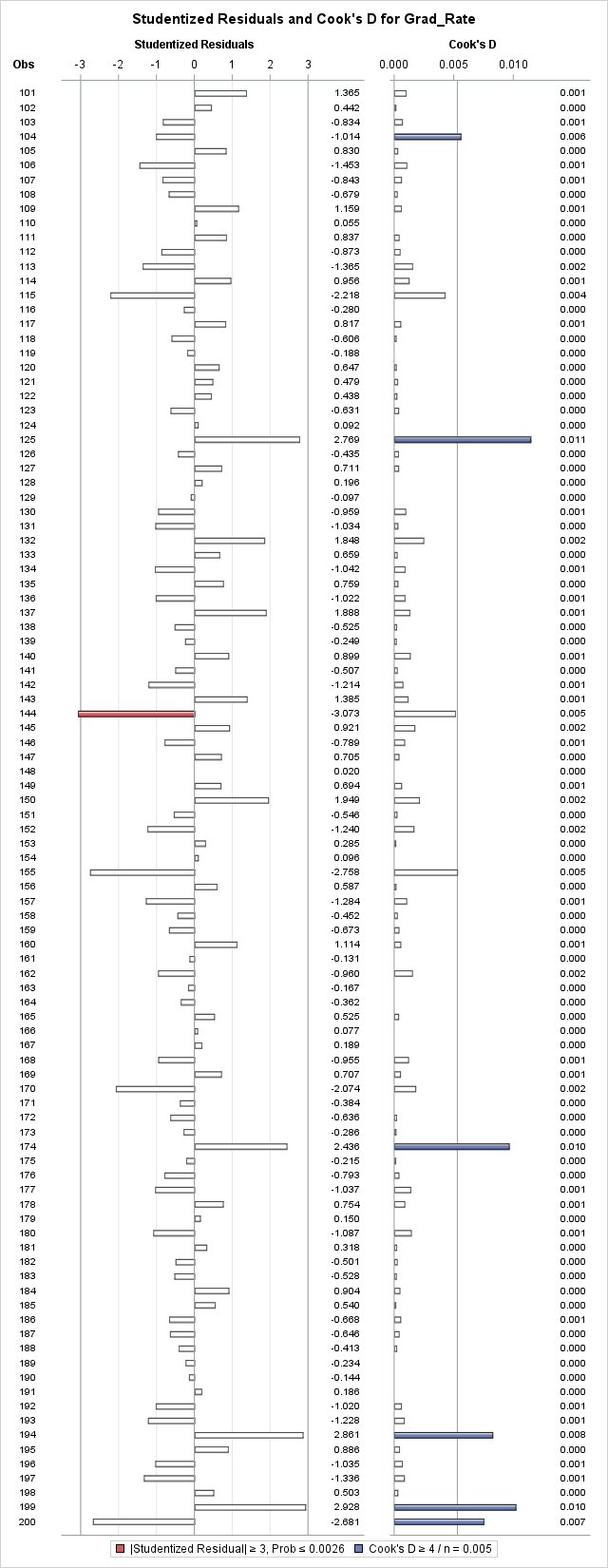
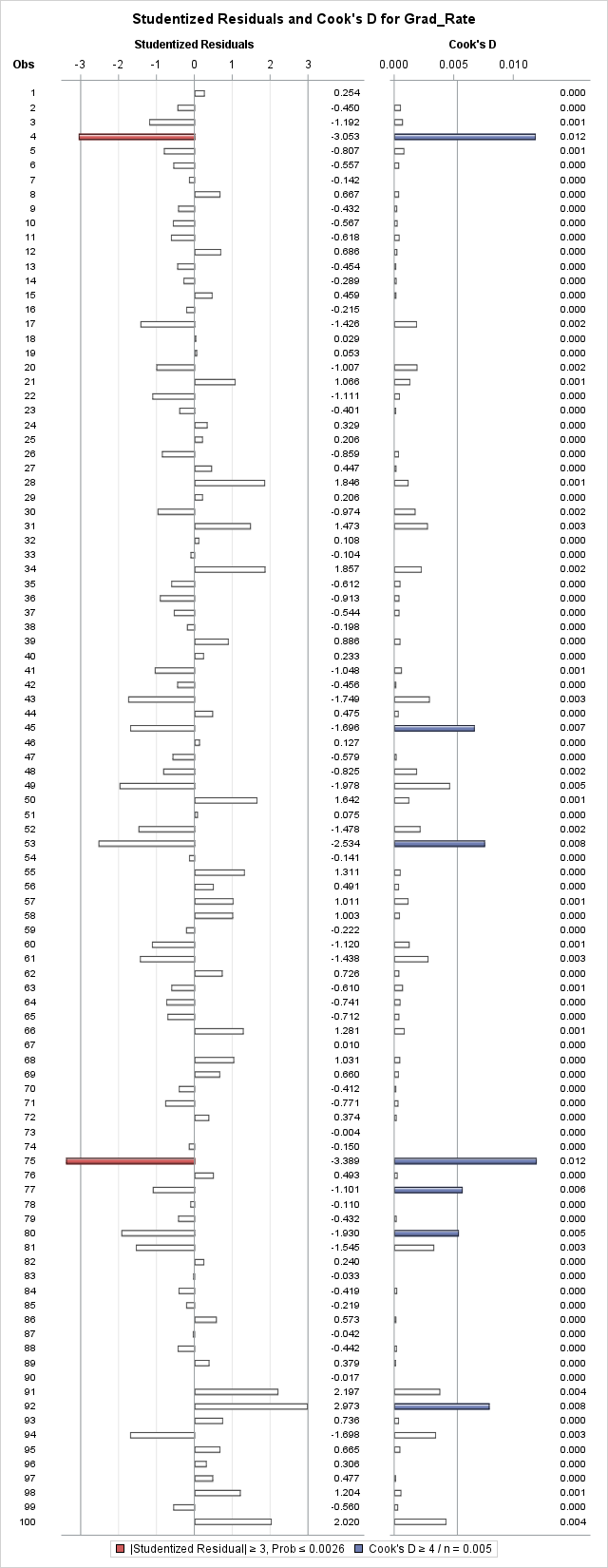
This graph indicates outliers within the data; with points upwards of 3 and downwards of -3. There is also a funnel pattern starting at 100 and opening up towards 50. This violated independence and constant variance.

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



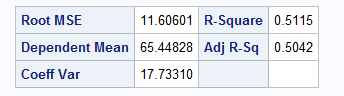
The normal probability plot represents a semi-straight line, making it normal; Normality is satisfied.

j) 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?



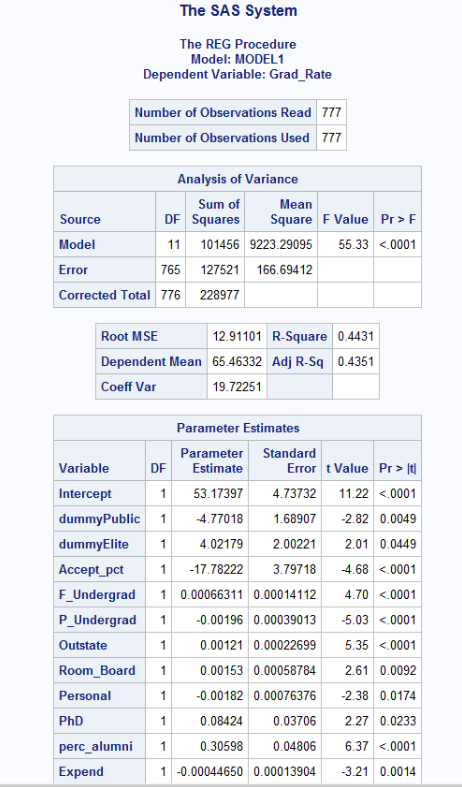
These graphs represent the final result - for all graphs please run the code. In total there were 23 outliers that were removed from the data set. These were all represented by arrows.

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



Based on the Adjusted R-Squared value of .5042 or 50.42%, about half of the variation in the graduation rate is explained by the model. Leaving the other half unexplained and a moderate association.

l) 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.

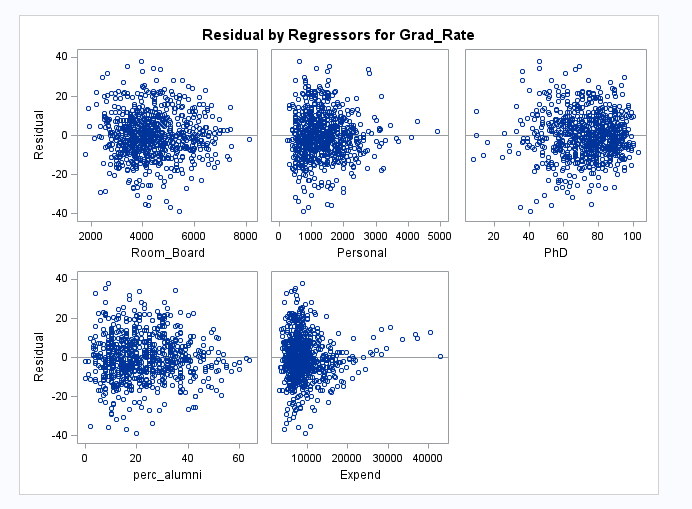
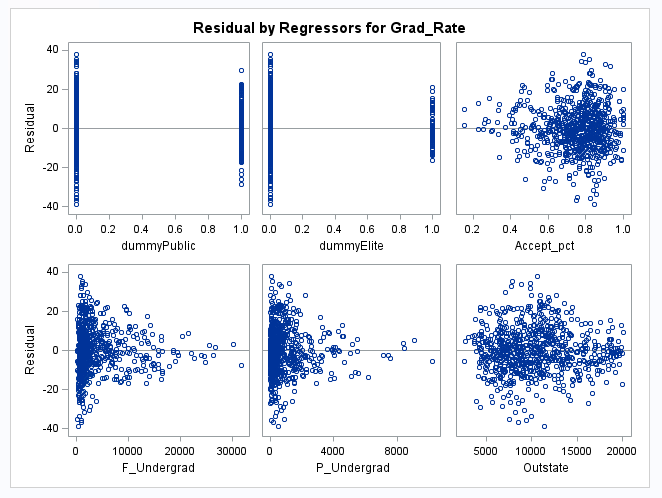


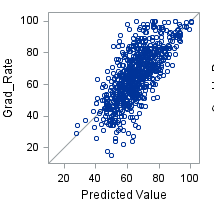
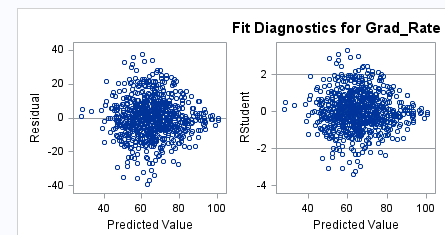
Looking at this chart - the biggest predictors are F\_Undergrad, Outstate, and perc\_alumni. In other words, those three predictors have the greatest influence on graduation rate. Also based on the box plot in c) elite universities have a higher graduation rate compared to non-elite universities. Between public and private universities - public perform better however not significantly.

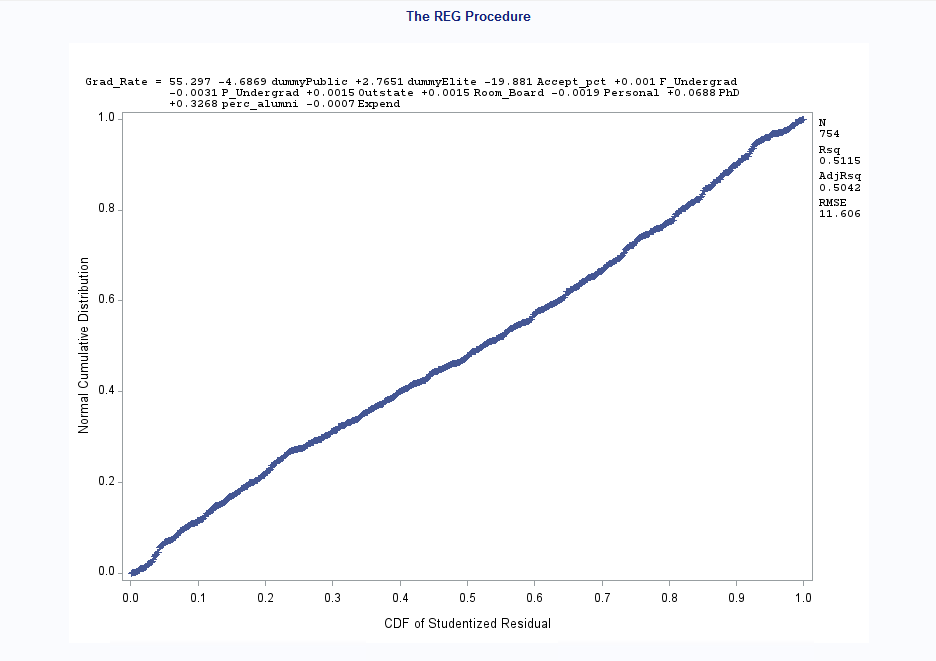
m) Use the final regression model to predict the graduation rate for the following values. Using SAS, compute the predicted graduation rate, 95% confidence interval and prediction interval for your estimate. Make sure to use SAS coding to determine the values. Include all relevant outputs. Discuss your findings.

| Private | Yes |
| --- | --- |
| Accept.pct | 0.87 |
| Elite10 | Not Elite |
| F.Undergrad | 3000 |
| P.Undergrad | 524 |
| Outstate | 6500 |
| Room.Board | 3300 |

| Books | 250 |
| --- | --- |
| Personal | 1350 |
| PhD | 40 |
| Terminal | 34 |
| S.F.Ratio | 30.2 |
| perc.alumni | 13 |
| Expend | 5201 |







Based on the above graphs - This university would have a slightly above mean for graduation rate for a private school and non-elite university.

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

proc import datafile="C:\Users\JDORETTI\Downloads\College.csv" out=College replace;

delimiter=',';

getnames=yes;

datarow=2;

run;

\*Histogram;

PROC UNIVARIATE normal;

VAR Grad\_Rate;

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

RUN;

\*Scatterplots;

PROC GPLOT;

PLOT Grad\_Rate\*(Accept\_Pct F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend Private Elite10);

RUN;

\*dummy variables;

data College2;

set College;

dummyPublic=(Private="No");

dummyElite=(Elite10="1");

RUN;

PROC SORT;

BY dummyPublic;

RUN;

\*Boxplot;

PROC BOXPLOT;

plot Grad\_Rate \* dummyPublic;

RUN;

PROC SORT;

BY dummyElite;

RUN;

\*Boxplot;

PROC BOXPLOT;

plot Grad\_Rate \* dummyElite;

RUN;

\*Regression model 1;

PROC REG;

MODEL Grad\_Rate=dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend;

RUN;

\*Regression model 2;

PROC REG;

MODEL Grad\_Rate=dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend;

RUN;

\*VIF regression;

PROC REG;

MODEL Grad\_Rate=dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend/VIF;

RUN;

\*Regression model - forward;

PROC REG;

MODEL Grad\_Rate=dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend/ selection=forward;

RUN;

\*Regression model - backward;

PROC REG;

MODEL Grad\_Rate=dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend/ selection=backward;

RUN;

\*VIF for regression model;

PROC REG;

MODEL Grad\_Rate=dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend/VIF;

RUN;

\*Residual Plot;

PLOT student.\*predicted.;

PLOT student.\*(dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend);

PLOT npp.\*student;

RUN;

\*Influential points and Outliers;

PROC REG;

MODEL Grad\_Rate=dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend/influence r;

PLOT student.\*(dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend predicted.);

PLOT npp. \*student.;

RUN;

\*Removed outliers;

data College3;

set College2;

if \_n\_ in (39, 57, 75, 88, 199, 238, 293, 360, 529, 563, 619, 640, 725, 730, 758) then delete;

RUN;

\*Outliers;

PROC REG;

MODEL Grad\_Rate=dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend/influence r;

PLOT student.\*(dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend predicted.);

PLOT npp. \*student.;

RUN;

\*Removed outliers;

data College4;

set College3;

if \_n\_ in (113, 234, 278, 464, 492, 573, 652, 742) then delete;

RUN;

\*Outliers;

PROC REG;

MODEL Grad\_Rate=dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend/influence r;

PLOT student.\*(dummyPublic dummyElite Accept\_pct F\_Undergrad P\_Undergrad Outstate Room\_Board Personal PhD perc\_alumni Expend predicted.);

PLOT npp. \*student.;

RUN;