**CSC423: Data Analysis And Regression / CSC 324: Data Analysis & Statistical Software II**

**Assignment-5** | **Total Points: 25 pts for CSC 423/25 pts for CSC 324**

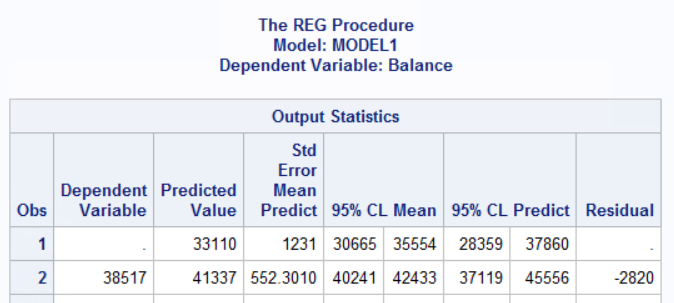
**Due Date: 02/20/2018 by 11:59 pm**

***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] – to be answered by everyone**

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.



Predicted balance value = 33,110 with CI (30665, 35554) and PI(28359, 37860)

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

Title "Analysis of Banking Data2";

**PROC** **IMPORT** datafile = 'S:\Homeworks\bankingfull.txt' out=bankingfull replace;

delimiter = '09'x;

getnames=YES;

datarow=**2**;

**run**;

**proc** **print**;

**run**;

\*Model 2, remove homeval to start;

Title 'Model 2';

**proc** **reg**;

model balance = age education income wealth/vif influence r stb;

**run**;

\* compute predictions on new value;

title "Compute Predictions";

\* STEP 6.1: creates dataset with new value;

**data** pred;

input age education income homeval wealth;

datalines;

34 13 89000 160000 140000

;

**run**;

**proc** **print**;

**run**;

\* STEP 6.2: join new dataset with bank dataset;

**data** prediction;

set pred bankingfull;

**run**;

**proc** **print**;

**run**;

**proc** **reg** data = prediction;

MODEL balance = age education income wealth/CLI CLM;

\*use cli for prediction intervals, and clm for intervals for averages;

**RUN**;

**PROBLEM 2 [20 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*

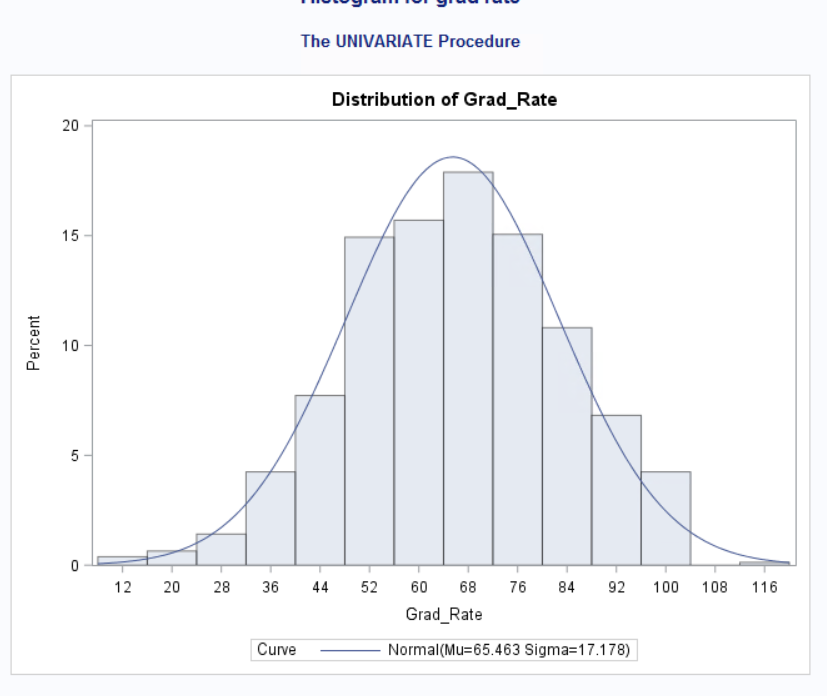
*Grad.Rate Graduation rate in 4 years*

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.

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).



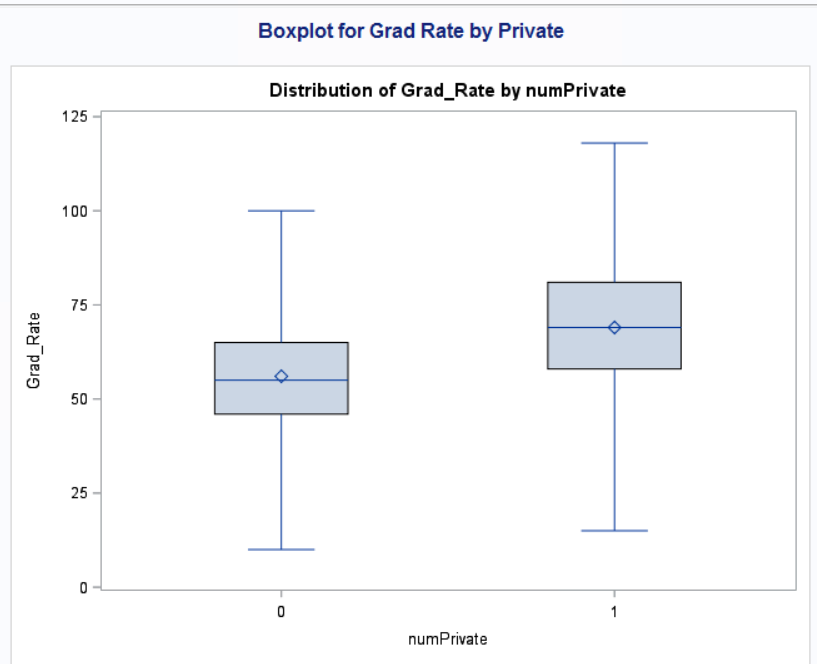
**This is a normal distribution.**

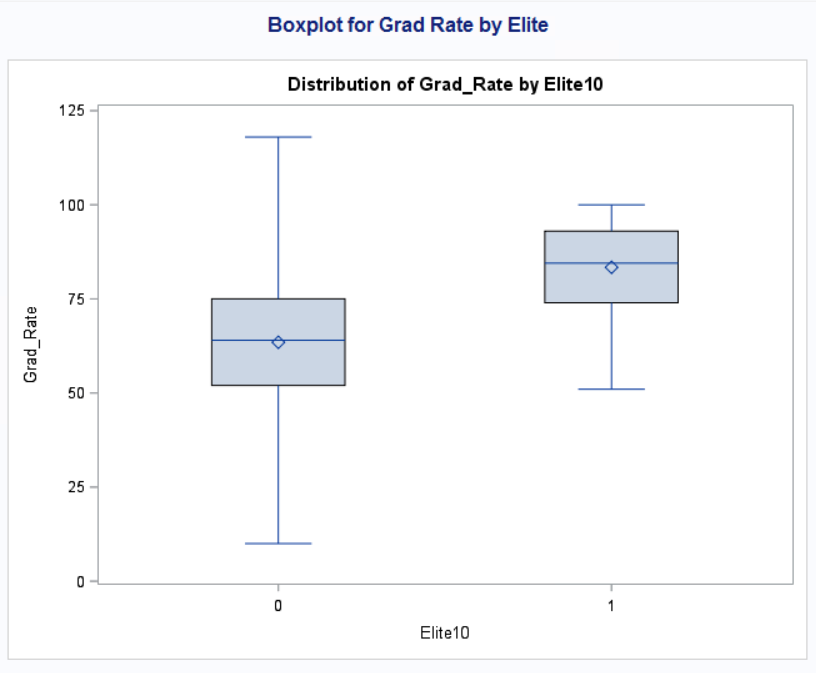
1. Create scatterplots for Grad.Rate vs each of the independent variables. What conclusions can you draw about the relationships between Grad.Rate and the independent variables? (No need to include the scatterplots in your submission).

**There is a positive correlation between grad\_rate and outstate, and grad rate and room\_board.**

**You cannot conclude any other relationships between grad\_rate and the other variables.**

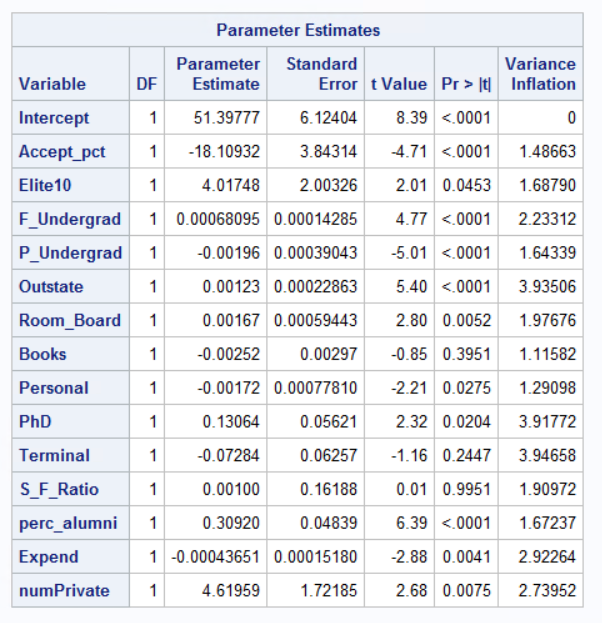
1. Build boxplots to evaluate if graduation rates vary by university type (private vs public) and by status (elite vs not elite). Include the boxplots and discuss your findings. (See SAS Procedures section on D2L if you need the code to generate a boxplot).

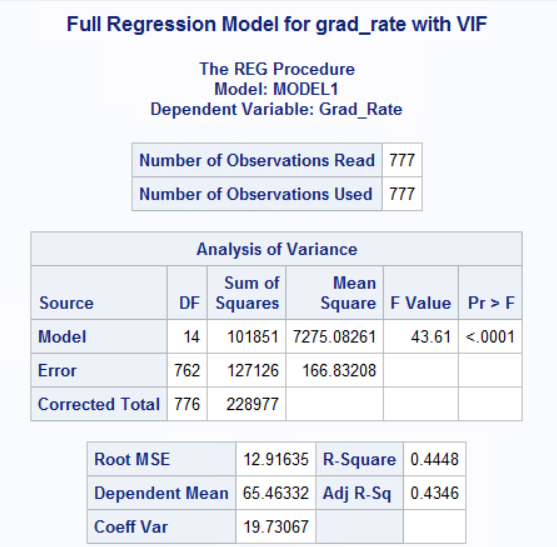




1. Fit a full model (with all independent variables) to predict Grad.Rate. Discuss the parameter estimates, significance, goodness-of-fit and AdjR2 values. Include the relevant output.

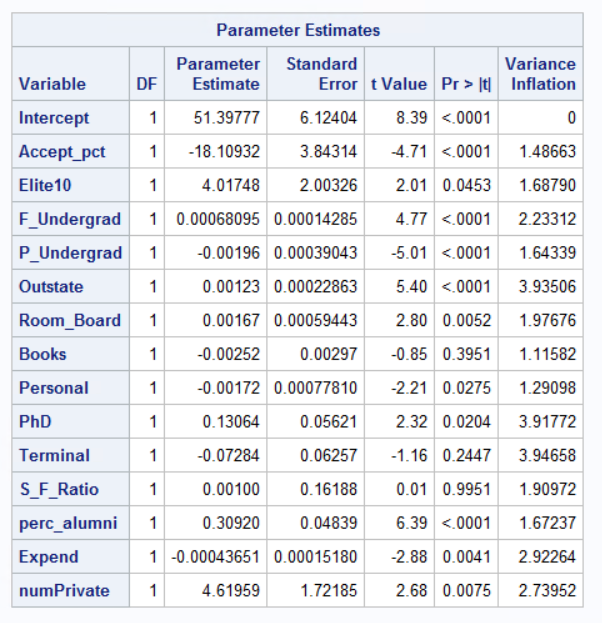
**According to the F-value, the model has a good fit. The AdjR2 = 0.4346 meaning 43% of the variance can be explained by the independent variables. Room\_Board, Books, Terminal, and S\_F\_Ratio have p-values <0.05 and should be excluded from the model.**

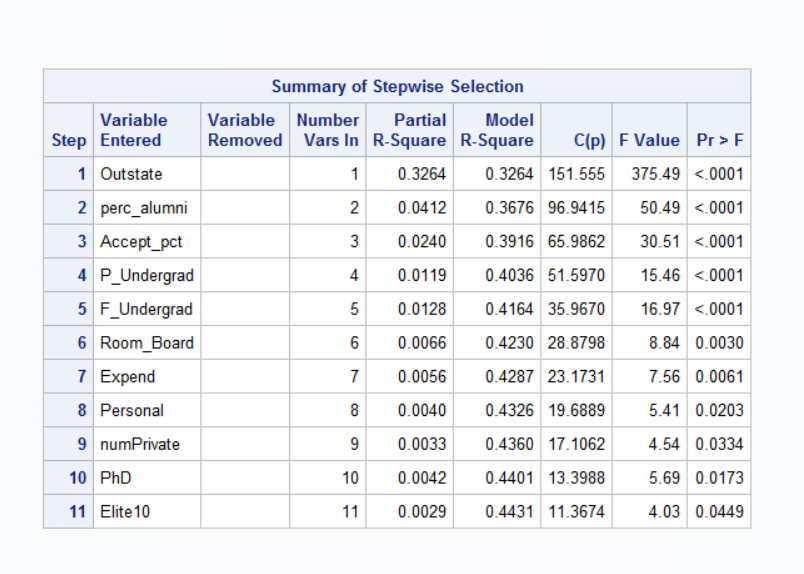


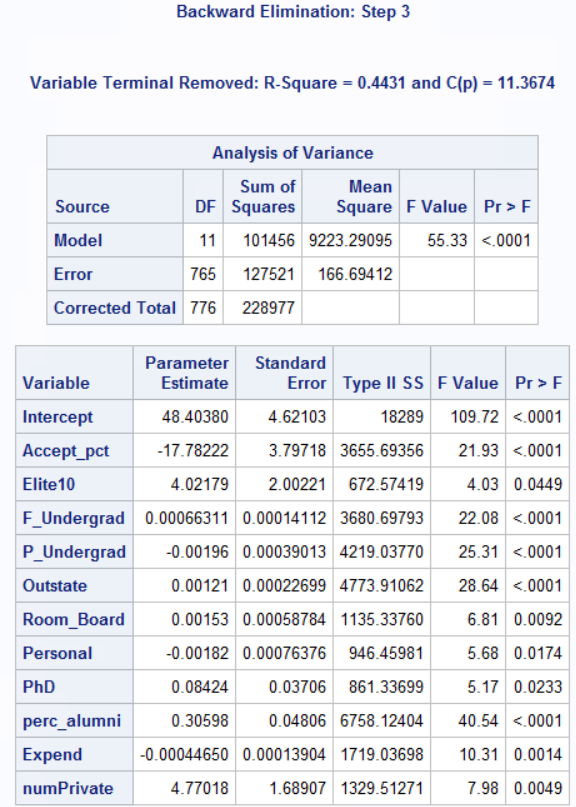


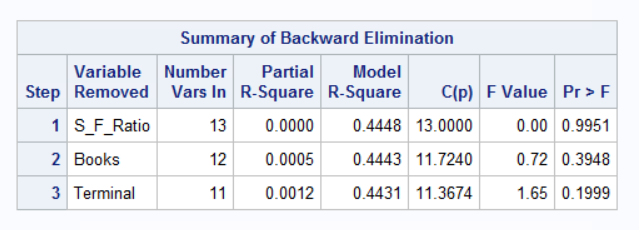
1. Does multi-collinearity seem to be a problem here? What is your evidence? Compute and analyze the VIF statistics. Include the relevant output and discuss your answer.

**Multi-collinearity does not appear to be a problem here because none of the values have a VIF >10.**

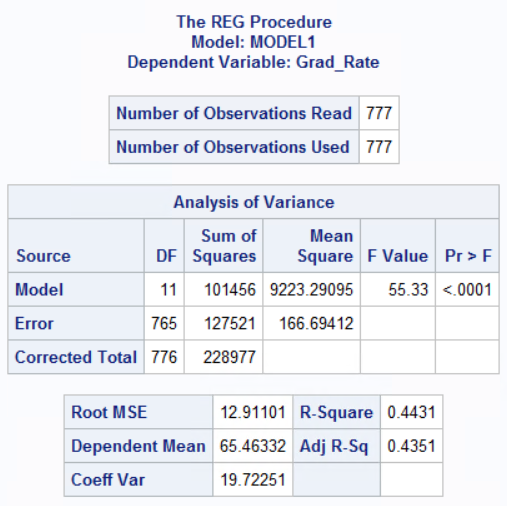


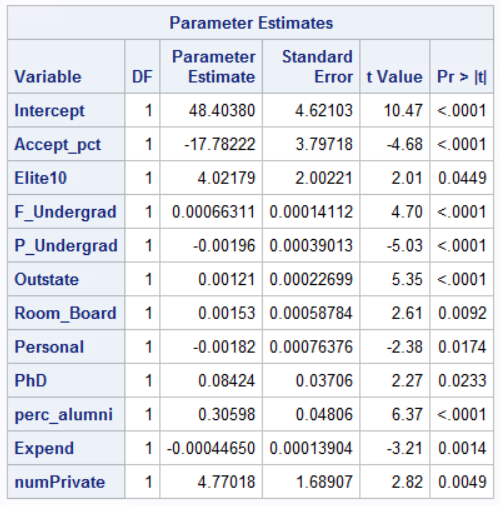
1. Apply TWO variable selection procedures to find an optimal subset of independent variables to predict Grad.Rate*.* You can choose any two procedures among the ones we learned in class: backward selection, forward selection, adj-R2, Cp, stepwise. Make sure to include the o/p of the 2 selection methods.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





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**.





**Grad\_rate = 48.4 -17.78Accept\_pct + Elite104.022 + 0.0006F\_Indergrad – 0.00196P\_undergrad + 0.00121Outstate + 0.00153Room\_Board – 0.00182Personal + 0.084PhD + 0.31Perc\_alumni -0.0004Expend +4.77numPrivate**

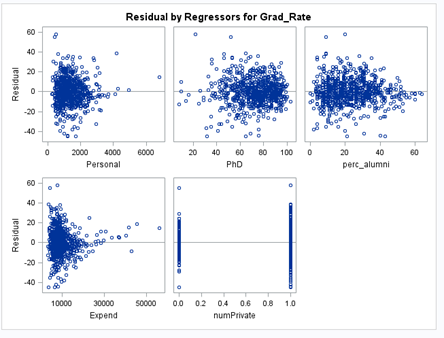
**Where numprivate =1 when private = ‘Yes’**

**I chose this model because it was the outcome of both the backward and stepwise model selections.**

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.

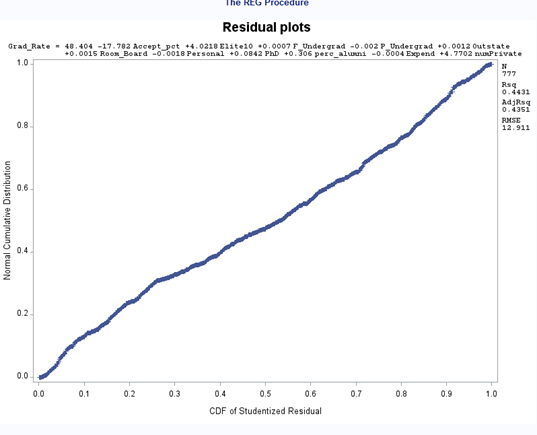
**The plot does not have independence or constant variance. It is also not linear.**





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

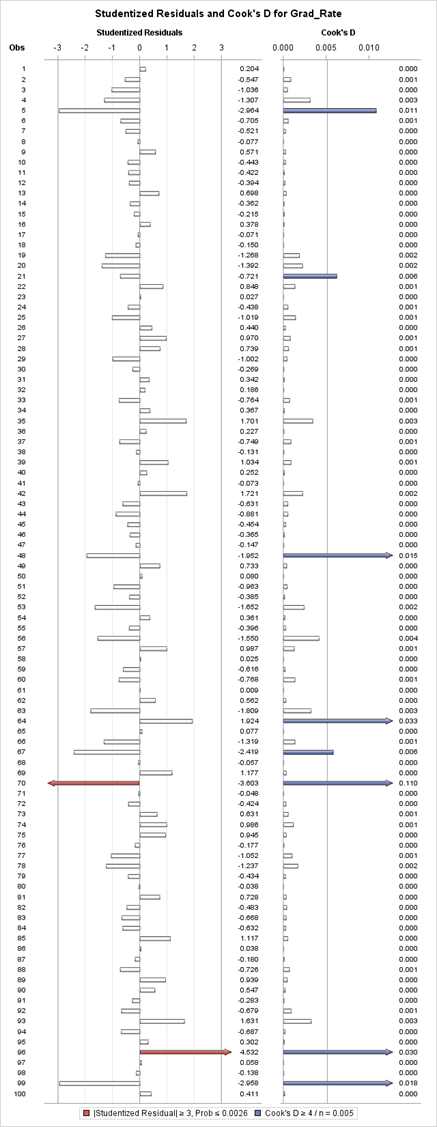
**The plot shows that the distribution is normalized.**

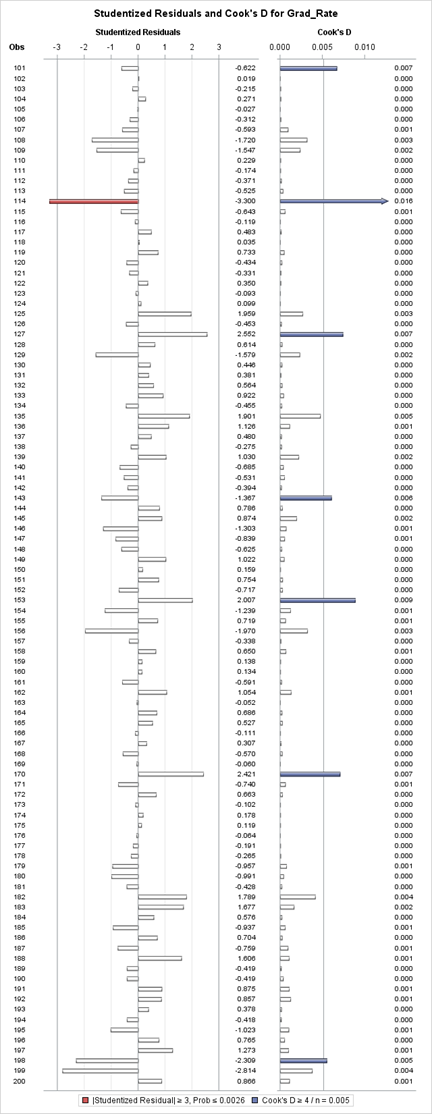


1. Are there any outliers or Influential Points? Compute appropriate statistics. Include the outputs and explain why/why not?

**Yes, there are both outliers and Influential points.**

**The following data points are both outliers and influential points: 70, 96, 114, 318. 378, 395, 586,**





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

**The AdjR2 = 0.4346 which means that 43% of the variability in graduation rates can be explained by the x-variables (accept\_pct, elite10, f\_undergrad, etc).**



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.

**Outstate is the most significant predictor in the model.**

**Private universities have a 4.6% increase in graduation rates as compared to public universities.**

**We know that there is a 4% increase in graduation rate when a university is elite vs non-elite.**

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

/\* import data from bankdata.txt file and save it in the bank sas dataset \*/

title "Import College Data";

**proc** **import** datafile="S:\Homeworks\college.csv" out=college replace;

getnames=yes;

**run**;

**data** college;

set college;

numPrivate = (Private = "Yes");

**run**;

**proc** **print**;

**run**;

\* analyze distribution of response variables;

title "Histogram for grad rate";

**proc** **univariate**;

var Grad\_Rate;

histogram/normal;

**run**;

\* create scatterplots;

title "GPLOTS for Y and X-variables";

**proc** **gplot** data=college;

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

**run**;

\*create matrix scatterplot;

title "Scatterplot Matrix";

**proc** **sgscatter** data=college;

matrix Grad\_Rate Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend numPrivate;

**run**;

\*create boxplot;

title "Boxplot for Grad Rate\*Private";

**proc** **sort** data=college;

by numPrivate;

**proc** **boxplot** data=college;

plot gran\_rate\*numPrivate;

**run**;

\*create boxplot;

title "Boxplot for Grad Rate\*Elite10";

**proc** **sort** data=college;

by Elite10;

**proc** **boxplot** data=college;

plot gran\_rate\*elite10;

**run**;

title "Full Regression Model for grad\_rate with VIF";

**proc** **reg** data=college;

model Grad\_Rate =Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend numPrivate/vif;

**run**;

title "Selection Method-2: backward Selection Method";

\* Selection Method-2: backward Selection Method;

**proc** **reg** data=college;

model Grad\_Rate = Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend numPrivate/selection = backward;

**run**;

title "Selection Method-3: stepwise Selection Method";

\* Selection Method-3: stepwise Selection Method;

**proc** **reg** data=college;

model Grad\_Rate = Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend numPrivate/selection = stepwise;

**run**;

Title 'Residual plots';

plot student.\*predicted;

plot student.\*(Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend numPrivate);

plot npp.\*student.;

**run**;

\* final model analysis with options r, influence, vif, stb;

title "Final Model w/ options";

**proc** **reg**;

model Grad\_Rate = Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend numPrivate/vif r influence stb;

**run**;