**Aamir patel**

**Data Analysis And Regression**

**Assignment-4** | **Total Points: 26**

**This assignment is longer to make up for the midterm exam. Make sure to start the assignment the very next day.**

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 [16 pts]**

The file bankingfull.txt attached to this assignment contains the full dataset. You analyzed a smaller set for a previous assignment. It provides data acquired from banking and census records for different zip codes in the bank’s current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The data show

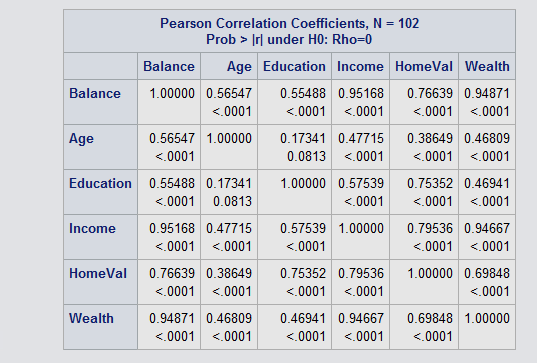
* median age of the population (AGE)
* median years of education (EDUCATION)
* median income (INCOME) in $
* median home value (HOMEVAL) in $
* median household wealth (WEALTH) in $
* average bank balance (BALANCE) in $

The goal of this exercise is to define a regression model to predict the average bank balance as a function of the other variables.

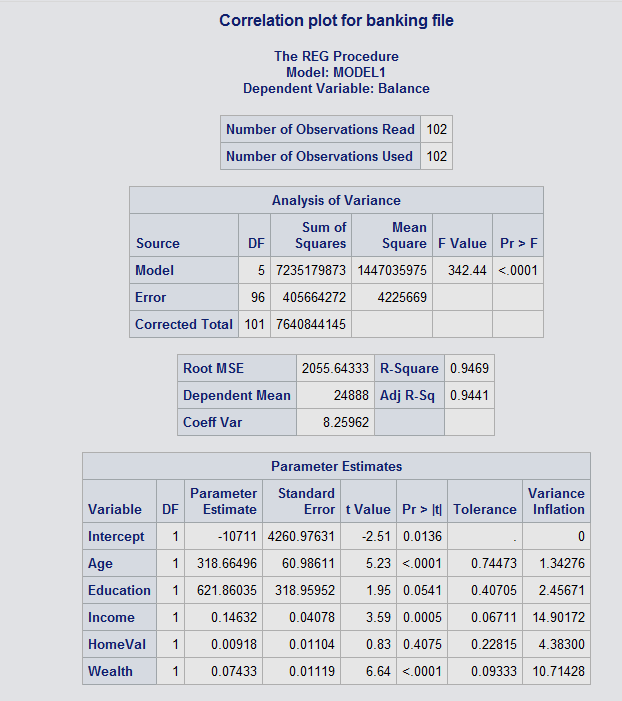
1. Create scatterplots to visualize the associations between bank balance and the other five variables. Include the relevant output. Discuss the patterns displayed by the scatterplot. Also, explain if the associations appear to be linear? (you can create either scatterplots or a matrix plot)

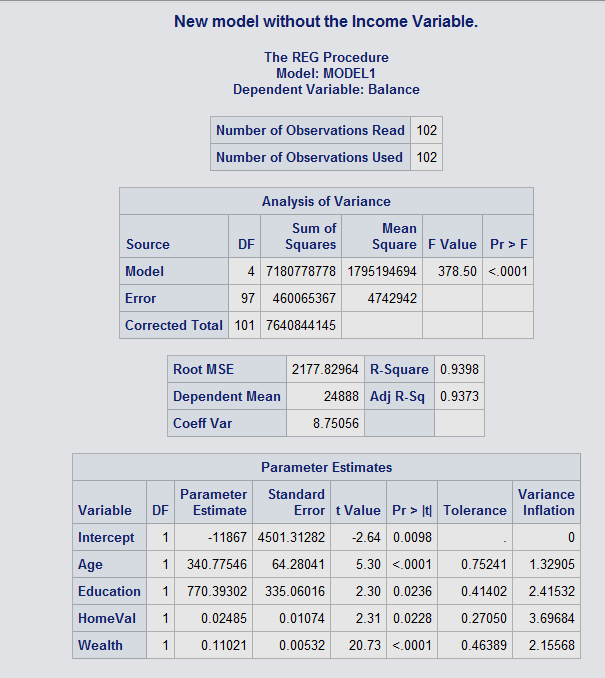
* 
* Balance with Age: - This scatterplot shows a weak, positive, linear association between Balance and Age. There appear to be no obvious outliers that appear in this matrix scatter plot.
* Balance with Education – This scatterplot shows a weak, positive, linear association between Balance and Education. There appears to be a couple of outliers on the x side of the graph.
* Balance and Income: - This Scatterplot has a strong positive linear association between Balance and Income. There appears to be no outliers in this scatter plot.
* Balance and HomeVal – This scatterplot has a strong positive linear association between Balance and HomeVal. There appears to be no outliers in this scatterplot.
* Balance and Wealth: - This scatterplot has a strong positive linear association between Balance and Wealth. There appears to be no outliers in this scatterplot.

1. Compute correlation values of bank balance vs the other variables. Include the relevant output. Interpret the correlation values, and discuss which variables appear to be strongly associated.

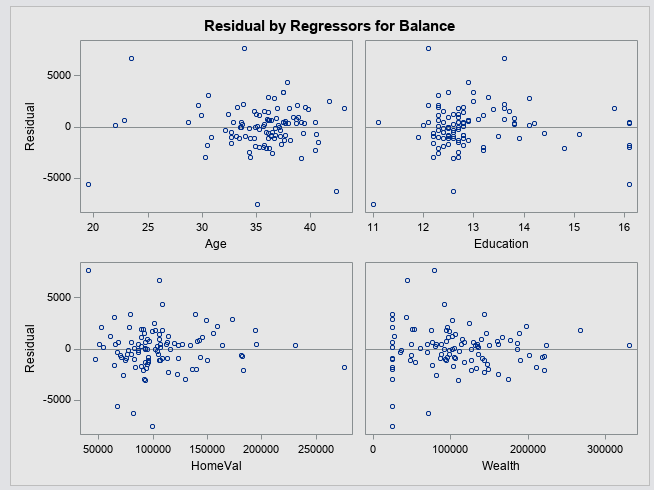
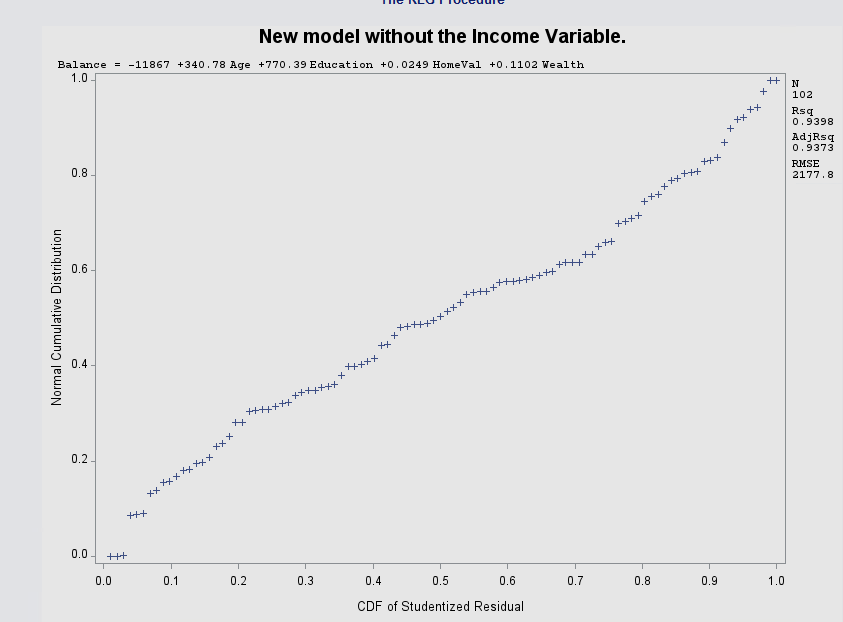
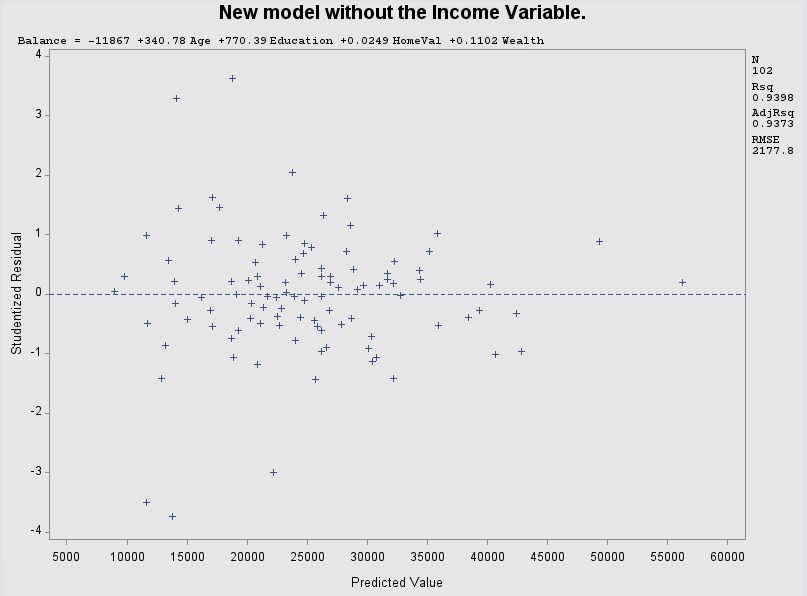
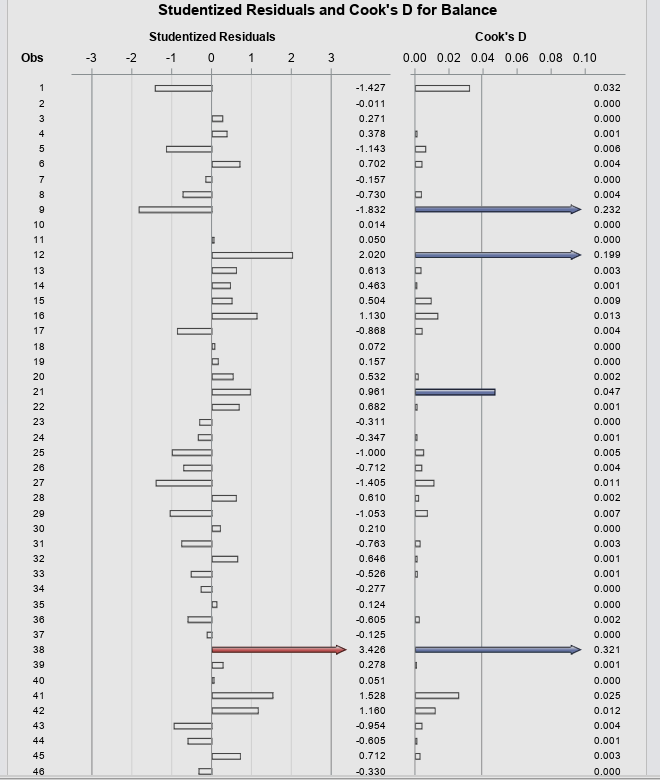
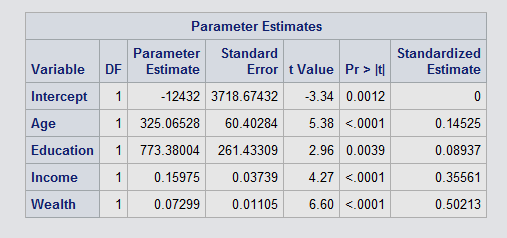
* 
* As expected Income, HomeVal, and Wealth seems to be very much correlated with Balance. Each of them has a very high correlation. While Age and Education seem to have a weak correlation with balance.
* Balance and income has a correlation of 0.95. (which is positively correlated and is also strong).
* Balance and HomeVal has a medium correlation value of 0.76. (which is positively correlated but not as strong as it should be.)
* Balance and Wealth has a high correlation value of 0.94. (which is positively correlated and is also strong).
* Balance and Age has a low correlation of 0.46. (which is positively correlated but not strong).
* Balance and Education has a low correlation of 0.46. (which is positively correlated but not strong).

1. Fit a regression model of balance vs the other five variables (model M1). Compute the VIF statistics for each x-variable and analyze whether there is a problem of multicollinearity and take appropriate action. Include the relevant output. Discuss your answer.



* Upon running and analyzing the regression model we can see that there are two variables with the VIF > 10. Income at 14.9 and Wealth at 10.7. So in order to fix the multi collinearity we can need to remove the varible with highest VIF and re run the model. So, once we re run the model we will get this: -
* 
* Upon removing the Income variable, we can see that it has fixed the multicollinearity issue that we had. So, we can keep this model now.

1. Apply your knowledge of regression analysis to define a better model M2. Include the SAS output for both models and answer the following questions :
   1. Analyze the adj-R2 values for both models M1 and M2. Which model has the largest adj-R2 value?

* The adj - R^2 for model 1 is : - 0.94 \*100 = 94%.
* The adj - R^2 for model 2 is : - 0.93 \* 100 = 93%.
* We can see that the model 1 is has a higher adj - R^2.
  1. Create residual plots for M2 (Studentized residuals vs predicted; Studentized residuals vs x-variables; and normal plot of residuals). Analyze the residual plots to check if the regression model assumptions are met by the data. Include the relevant output and discuss your analysis.
* 
* 
* 
* Upon analyzing the residual model for student, we can see that studentized vs predicted model we can see that it is independent and does not follow a pattern.
* We can also see that NPP chart is very linear which indicates that it is a good model.
  1. Analyze if there are any outliers and/or influential points for your M2 model. If so, what actions would you take to address this issue? Make sure to implement any actions you specify here. Include the relevant output.
* 
* Upon analyzing the cook’D chart for studentized residuals we can see that observation 38, 9 and 12 is flagged, and upon a detailed observation we can tell that they are not outliers rather influentiatl points, so we should remove these values one at a time and rerun the program, starting with 38 and moving forward to 9 then 12.
  1. Compute the standardized coefficients for M2 and discuss which predictor has the strongest influence on balance? Include the relevant output.
* 
* We can see that wealth has the strongest influence on balance at 0.50. followed by Income at 0.35, Age at 0.145, and Education at 0.089.

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

\*Import Statement;

**PROC** **IMPORT** datafile="Bankingfull.txt" out=main\_data replace;

delimiter='09'x;

getnames=YES;

datarow=**2**;

**run**;

\*Print the entire database;

**proc** **print**;

**run**;

\*print scatterplots;

**proc** **sgscatter**;

title"Scatter matrix plot for banking file";

matrix Balance Age Education Income HomeVal Wealth;

**run**;

\*Print correlation;

**proc** **corr**;

title"Correlation plot for banking file";

var Balance Age Education Income HomeVal Wealth;

**run**;

\*Print Regression model.;

**proc** **reg**;

model Balance = Age Education Income HomeVal Wealth / vif tol;

**run**;

\*Re run the model without the highest VIF value.;

**proc** **reg**;

title"New model without the Income Variable.";

model Balance = Age Education HomeVal Wealth / vif tol;

**run**;

\* Residual model;

**PROc** **reg** CORR;

model Balance = Age Education HomeVal Wealth;

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

plot student.\*predicted.;

plot npp.\*student.;

**RUN**;

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

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

\*Reduced model;

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

\*Residual analysis;

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

\*Normal probability plot of residuals;

plot npp.\* residual.;

**RUN**;

\*Deleting observation # 38 from the dataset and naming it BankingFull;

**DATA** Banking;

set Bankingfull;

if \_n\_ = **38** then delete;

**RUN**;

\*New model without 38th observation.;

**PROC** **REG** data=Banking;

title"New model without 38th observation";

\*Reduced model;

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

\*Residual analysis;

plot residual.(predicted. Age Education Income Wealth);

\*npp residual graph;

plot npp.\*residual.;

**RUN**;

**Problem 2 [10 pts]**

Analytics is used in many different sports and has become popular with the Money Ball movie. The pgatour2006.csv dataset contains data about 196 tour players in 2006. The variables in the dataset are:

* Player’s name
* PrizeMoney = average prize money per tournament

And a set of metrics that evaluate the quality of a player’s game.

* DrivingAccuracy = percent of times a player is able to hit the fairway with his tee shot
* GIR = percent of time a player was able to hit the green within two or less than par (Greens in Regulation)
* BirdieConversion = percentage of times a player makes a birdie or better after hitting the green in regulation
* PuttingAverage = putting performance on those holes where the green was hit in regulation.
* PuttsPerRound= average number of putts per round (shots played on the green)

You are asked to build a model for PrizeMoney using the remaining predictors, and to evaluate the relative importance of each different aspects of a player’s game on the average prize money.

**Note:** For the non-golfers in the class, you can refer to this page for an explanation of the terms:

<http://en.wikipedia.org/wiki/Glossary_of_golf>

**SAS Code to Import the data**

\*import data from file;

**proc** **import** datafile="pgatour2006.csv" out=PGATour replace;

delimiter=',';

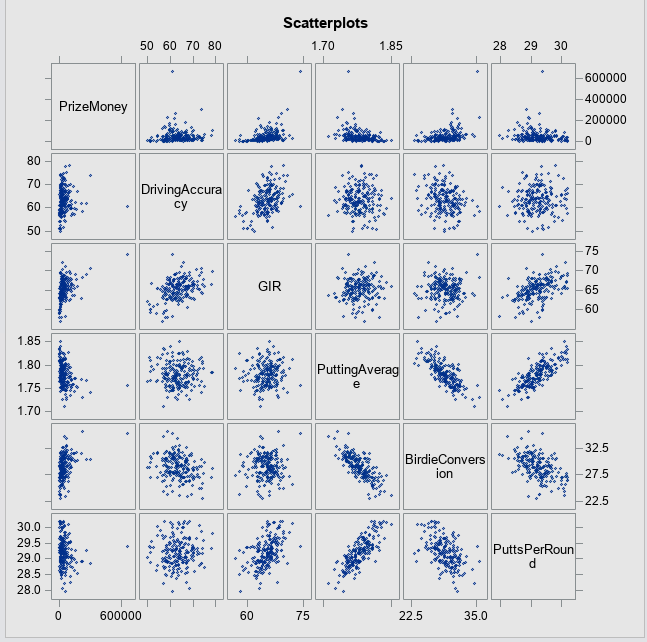
getnames=yes;

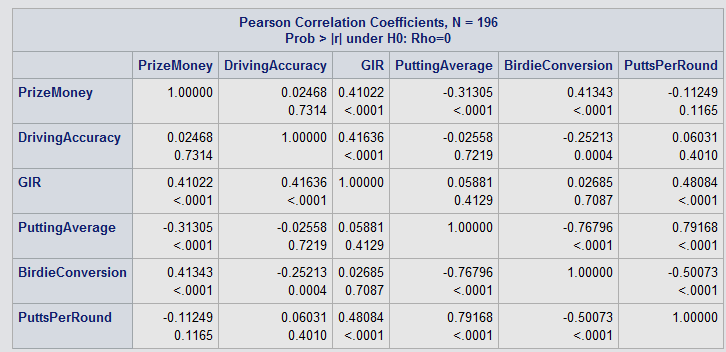
**run**;

Note:

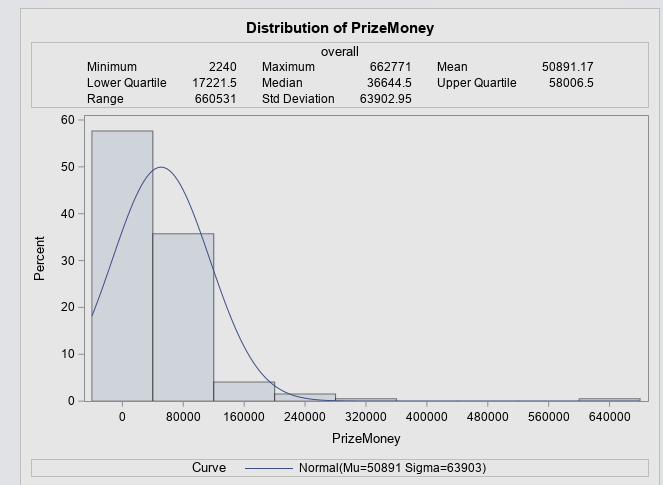
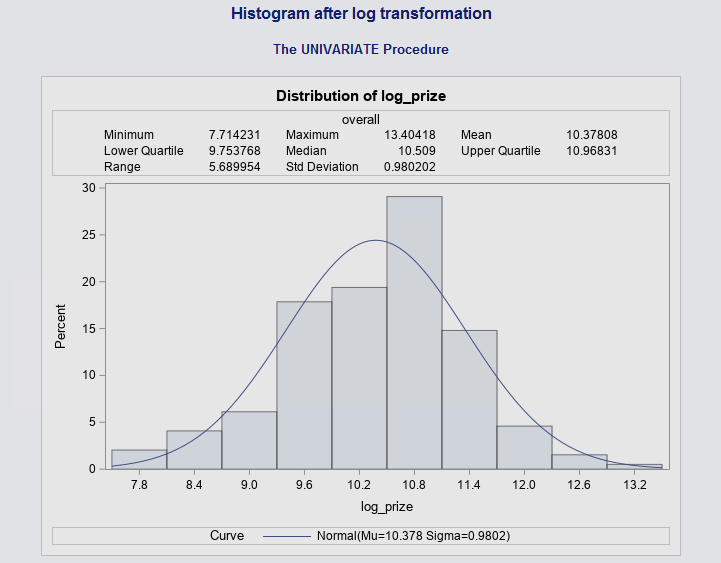
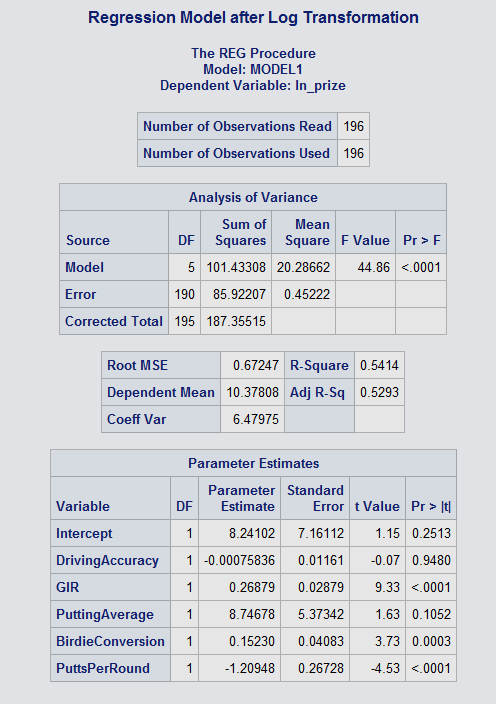
* The data file is in CSV format
* It is delimitered with a comma
* The SAS dataset it is writing into is PGATour. You can change the name if you like.

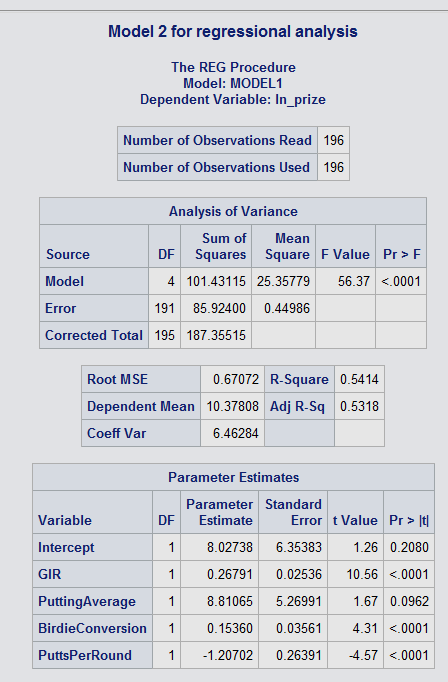
1. Create scatterplots to visualize the associations between PrizeMoney and the other 5 variables. Discuss the patterns displayed by the scatterplot. Also, explain if the associations appear to be linear? (you can create scatterplots or a matrix plot). Include the relevant output.



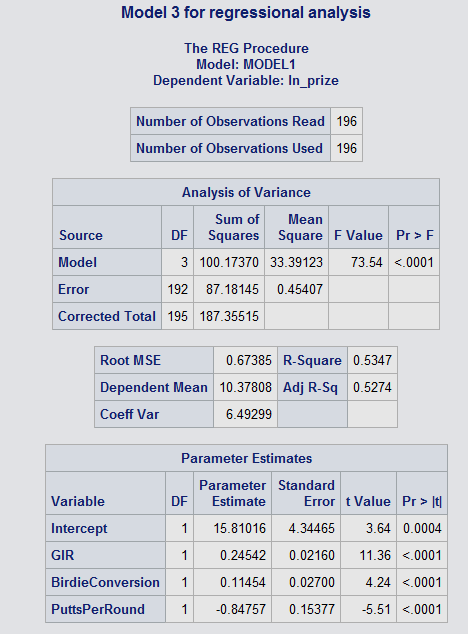


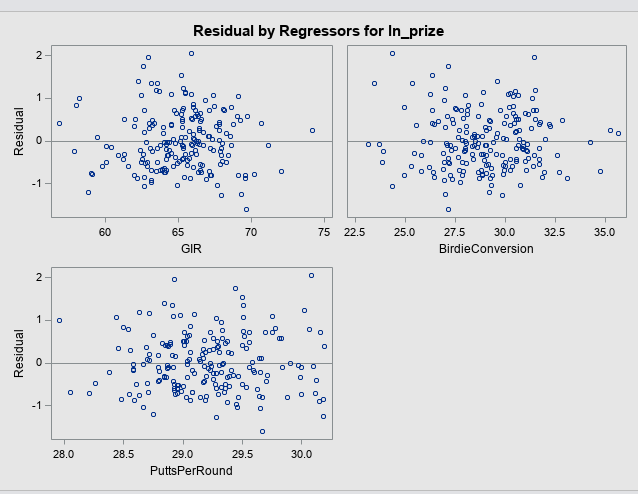
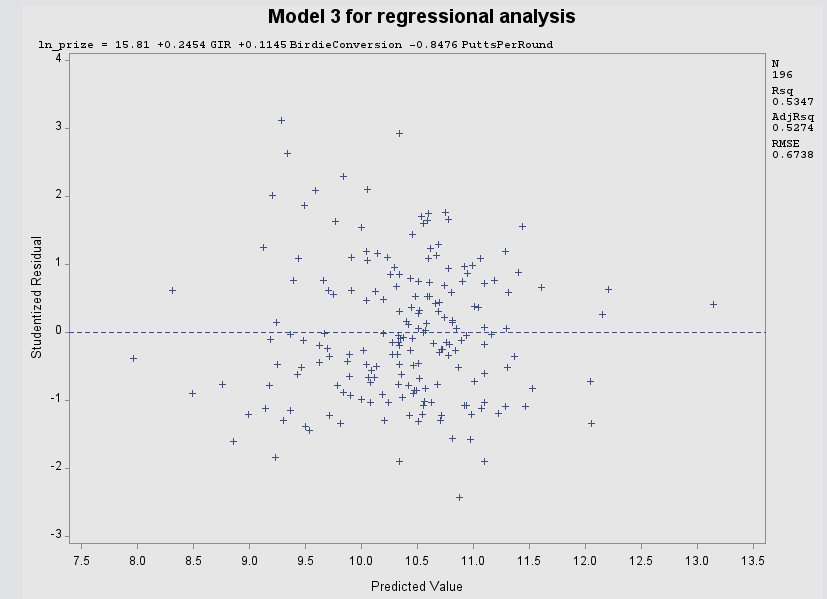
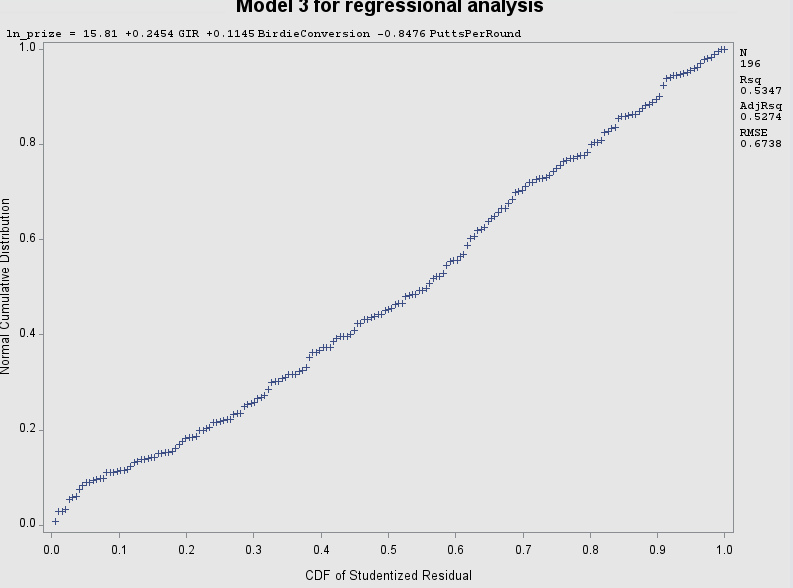
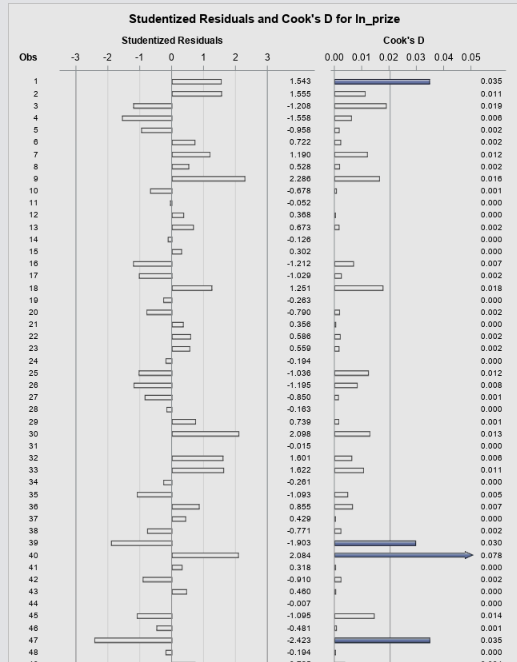
* + Upon looking at the Pearson Correlation Coefficient table, and scatterplots we can see that,
  + PrizeMoney and Driving Accuracy has a weak nonlinear positive association at 0.024.
  + PrizeMoney and GIR has a weak nonlinear positive association at 0.410.
  + PrizeMoney and PuttingAverage has a weak nonlinear negative association at -0.31.
  + PrizeMoney and Birdie Conversion has a weak nonlinear positive association at 0.413.
  + PrizeMoney and Putts Per Round has a weak nonlinear negative association at 0.11249.

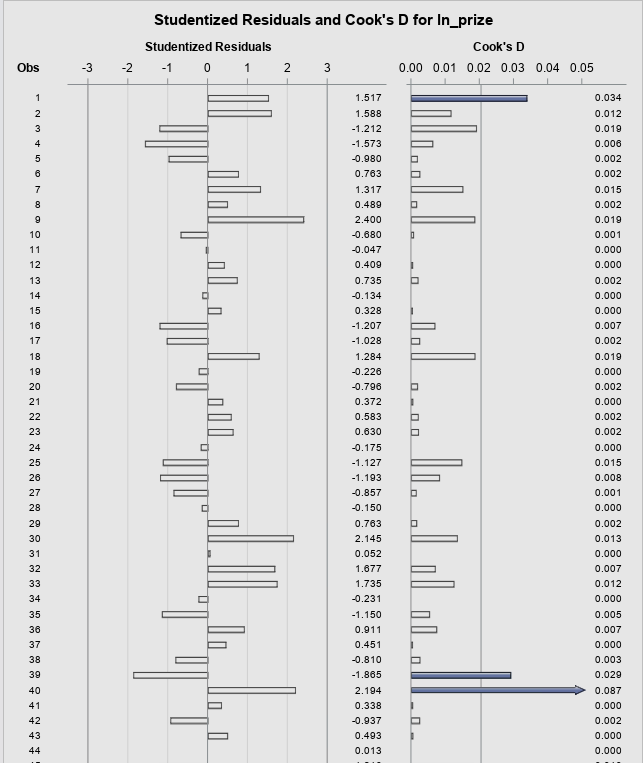
1. Analyze distribution of PrizeMoney, and discuss if the distribution is symmetric or skewed. Include the relevant output.
   * 
   * Upon analyzing this histogram we can clearly see that the data is very my skewed to the right, and we have a possible outlier at 640,000.
2. Apply a log transformation to PrizeMoney and compute the new variable ln\_Prize=log(PrizeMoney). Analyze distribution of ln\_Prize, and discuss if the distribution is symmetric or skewed. Include the relevant output.
   * 
   * Upon doing the log transformation we can see that the dataset is not skewing to the right anymore, and its a lot more symmetric now.
3. Fit a regression model of ln\_Prize using the remaining predictors in your dataset. Apply your knowledge of regression analysis to define a valid model to predict ln\_Prize. Include the outputs for all the questions below before you analyze them.
   * 1. If necessary remove the non-significant variables. Remember to remove one variable at a time (variable with largest p-value is removed first) and refit the model, until all variables are significant.
   * 
   * Upon creating the regression model, we can see that DrivingAccuracy is not significant in this model since its greater than 0.05 at 0.9480, so we can remove that. Followed by PuttingAverage at 0.1052.



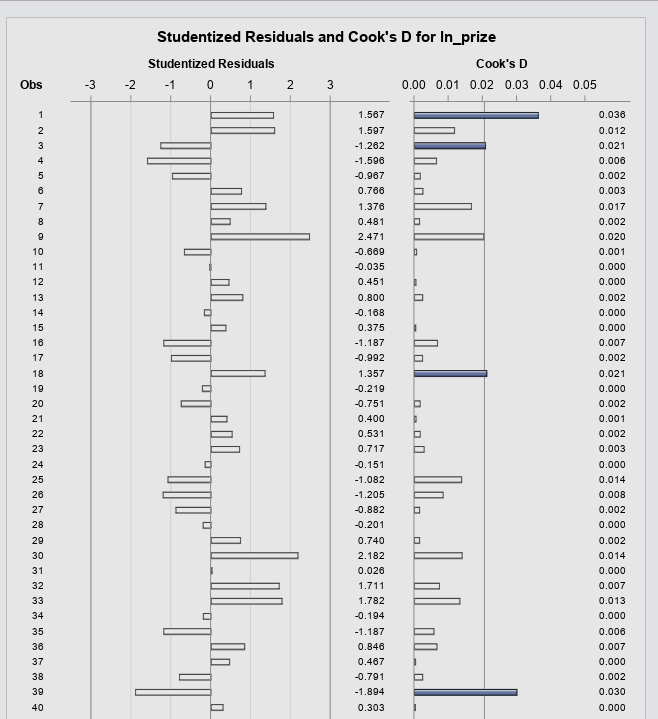
* + After removing the DrivingAccuracy model we can see that PuttingAverage is still not significant so we can remove that aswell.



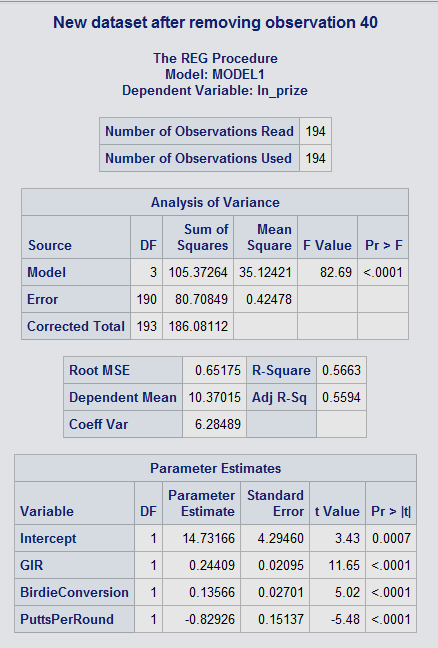
* + Now in the third model, we can confirm that all of the variables are significant enough to keep in our final model.
    1. Analyze residual plots to check if the regression model is valid for your data. Discuss your analysis.
  + 
  + 
  + 
  + Upon analyzing each of residual plots for each of our variables, we can see that the residual plot for GIR is fairly linear, independent and not following any patterns. We can see that BirdieConversion and PuttsPerRound is also very linear and independent.
  + The NPP plot also following the regression line almost perfectly, which indicates that it is a great model.
    1. Analyze if there are any outliers and/or influential points. If there are points in the dataset that need to be investigated, give one or more reason to support each point chosen. Take appropriate action(s) to implement it. Include the relevant outputs. Discuss your answer.
  + 
  + Upong analyzing the cook’s D for ln\_price we can see that it has marked observation # 185 as a possible outlier since their studentized residual rating is more than +3.
  + It has marked observations 185, 180, 141, 115, 101, 63, 60, 47, 39, 38, and 1 as possible influential points due to their cook’s D rating > 4/n or 1.
  + To possibly fix the issue, we need to start my removing observation 185, since it was marked for both influential point and outlier.



* + Upon removing observation 138, we can still see that it has maked observation 40,39, 60, 63, and 46 for possible influential points.
  + We can also notice removal of observation 138 has fixed our outlier issue.
  + Now lets remove observation 40, and re run.



Now this dataset is usable and we can continue with our analysis.

* + 1. Write down the final model equation. Discuss why this is the best model. Include all relevant statistics/values to substantiate your answer.
  + 
  + With our final model, we can now create our final model equation, which is.

Ln\_prize = 14.73166 + 0.24409 (GIR) + 0.13566 (BirdieConversion) + (-0.82926) (Putts Per Round)

* + This is the best model since its ajd-R^2 is highest of all 3 models at 55.9%.

1. Interpret the regression coefficients in the final model to answer the following question: How does an increase in 1% for GIR affect the average Prize money?

Model: -

Ln\_prize = 14.73166 + 0.24409 (GIR) + 0.13566 (BirdieConversion) + (-0.82926) (Putts Per Round)

* + So if we increase GIR by 1%, The Equation will be.
  + prize = 249,964,580.03 + 27.6459 + 14.52 + (-56.36)
  + So if we increase the GIR by 1%, the prize will increase by 27.64$.

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

\*import data from file;

**proc** **import** datafile="pgatour2006.csv" out=PGATour replace;

delimiter=',';

getnames=yes;

**run**;

title "main Dataset";

**proc** **print** ;

**run**;

title "Scatterplots";

**proc** **sgscatter** ;

Matrix PrizeMoney DrivingAccuracy GIR PuttingAverage BirdieConversion PuttsPerRound;

**RUN**;

title "Pearson Correlation";

**proc** **corr** ;

var PrizeMoney DrivingAccuracy GIR PuttingAverage BirdieConversion PuttsPerRound;

**RUN**;

\*Histogram;

title "Histogram";

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

var PrizeMoney;

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

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

pos=tm;

**run**;

\*Log transformation;

**data** PGATour;

SET PGATour;

ln\_prize=log(PrizeMoney);

**run**;

\*printing entire dataset to make sure log worked;

title"Dataset After Log Transformation";

**proc** **print**;

**run**;

\*Histogram after the log transformation;

title "Histogram after log transformation";

**PROC** **UNIVARIATE** normal data=PGATour;

var ln\_prize;

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

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

pos=tm;

**run**;

\*Regression Model after Log Transformation;

title"Regression Model after Log Transformation";

**proc** **reg** data=PGATour;

model ln\_prize = DrivingAccuracy GIR PuttingAverage BirdieConversion PuttsPerRound;

**run**;

title"Model 2 for regressional analysis";

**proc** **reg** data=PGATour;

model ln\_prize = GIR PuttingAverage BirdieConversion PuttsPerRound;

**run**;

title"Model 3 for regressional analysis";

**proc** **reg** data=PGATour;

model ln\_prize = GIR BirdieConversion PuttsPerRound;

plot student.\*(predicted. GIR BirdieConversion PuttsPerRound);

plot npp.\*student.;

**run**;

title"Analyzing for outliers and Influential points";

**proc** **reg**;

Model ln\_prize = GIR BirdieConversion PuttsPerRound / influence r;

**run**;

title"Removing obeservation 185";

**data** PGATourNew;

set PGATour;

if \_n\_ = **185** then delete;

**run**;

title"New dataset after removing observation 185";

**proc** **reg** data=PGATourNew;

Model ln\_prize = GIR BirdieConversion PuttsPerRound / influence r;

**run**;

title"Removing obeservation 40";

**data** PGATourNew2;

set PGATourNew;

if \_n\_ = **40** then delete;

**run**;

title"New dataset after removing observation 40";

**proc** **reg** data=PGATourNew2;

Model ln\_prize = GIR BirdieConversion PuttsPerRound / influence r;

**run**;