**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]**

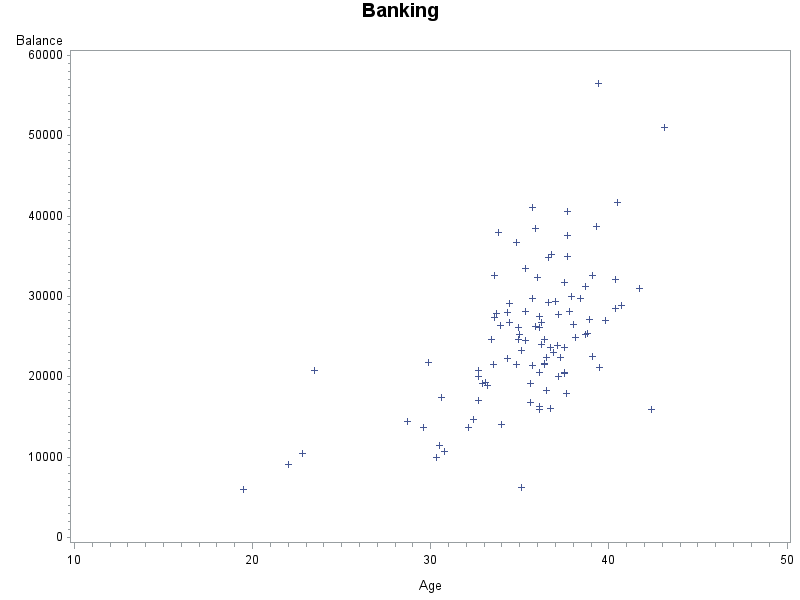
The file bankingfull.txt attached to this assignment contains the full dataset. You analyzed a smaller set for a previous assignment. It provides data acquired from banking and census records for different zip codes in the bank’s current market. Such information can be useful in targeting advertising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

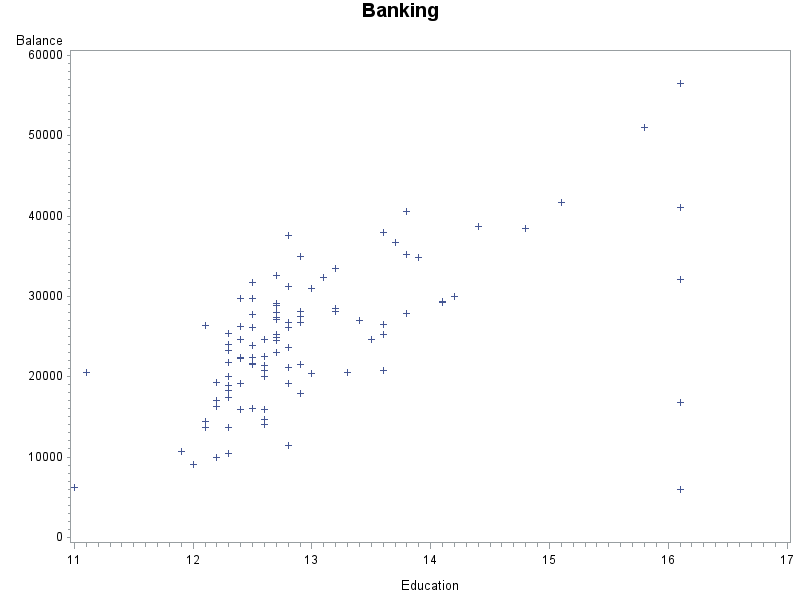
The goal of this exercise is to define a regression model to predict the average bank balance as a function of the other variables.

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)

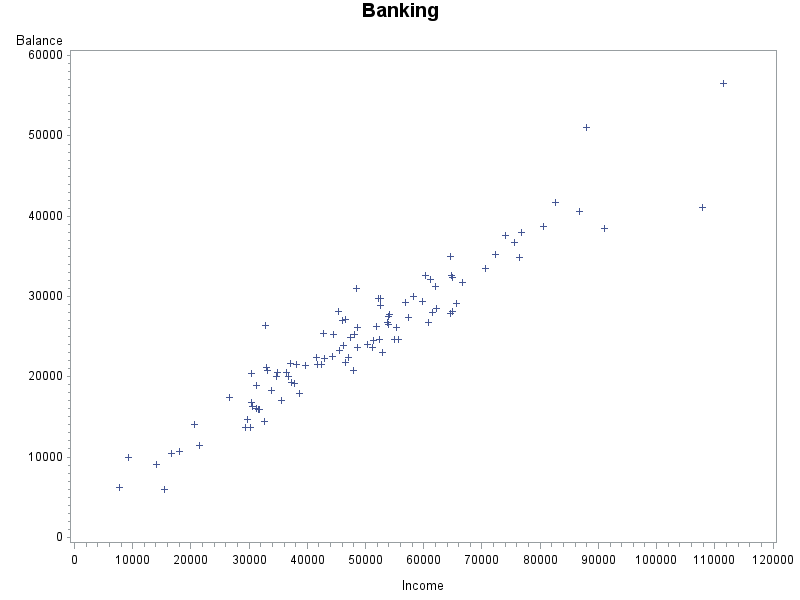
**Scatterplot for Balance and Age:**



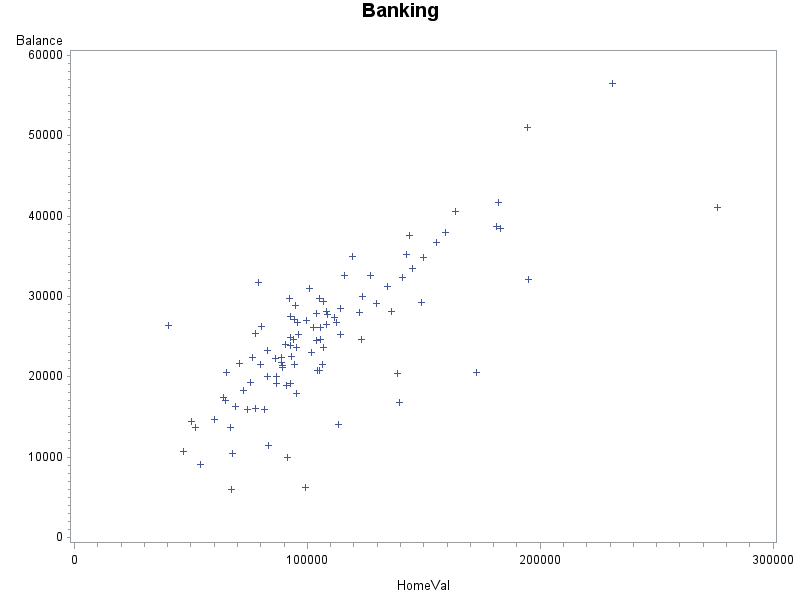
**Scatterplot for Balance and Education:**



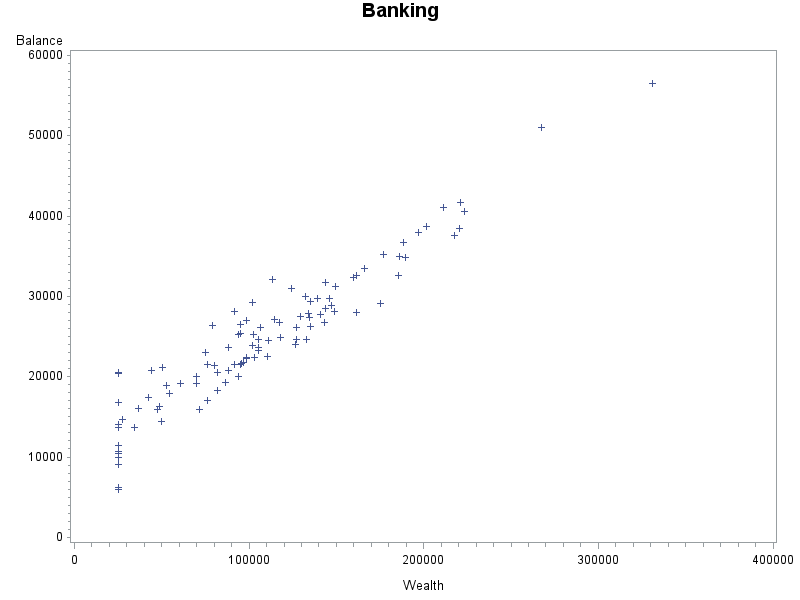
**Scatterplot for Balance and Income:**



**Scatterplot for Balance and HomeVal:**



**Scatterplot for Balance and Wealth:**



**My findings on the scatterplots**:

The first graph is a low positive correlation there is no really close data points where I can make out saying there is line created by the data to be high or perfect correlation. This is a linear positive scatterplot but data points are very spread out making it a low positive correlation.

The second graph is a low positive correlation the data is rather spread but there is a low positive correlation. I look at the whole x-axis at 16.1 where data points are plotted vertically but away from the rest of the majority of the data is 11.9 to 15 in education that is plotted but spread as well making it a low positive correlation.

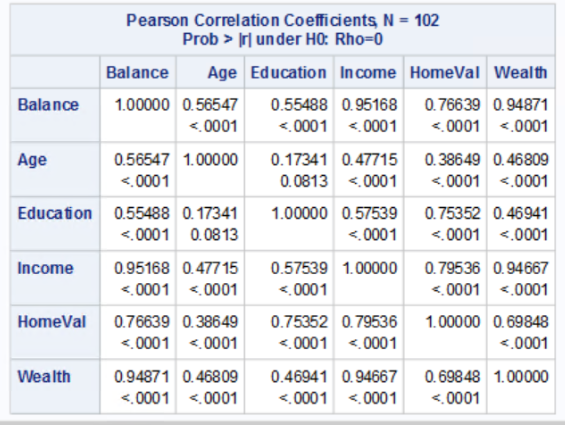
The third graph is a high positive correlation besides a couple distant points from the rest of the data from $8,000-$84,000 the data is very close and makes a linear line which to me is strong. It is a high positive correlation where the data is very close in spread and makes a clear linear line.

The fourth graph is a low positive correlation this spread is not totally far away it is like a moderate but what I say as considered to me as low because it is not close to high positive correlation. The data is spread out and from what I make as a line there is data underneath this line I see in my head which makes me consider it as low.

The fifth graph to me is a linear high positive correlation. From what I see the main data is in a linear fashion and the spread is very close together. I recognize that there is some distant points away from the majority of this line of data but that is why it is not perfect positive correlation. It is high with the points still being in a line with a few that might be a little outside but still close enough where it can be seen as high positive correlation.

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.

**Correlation Values:**



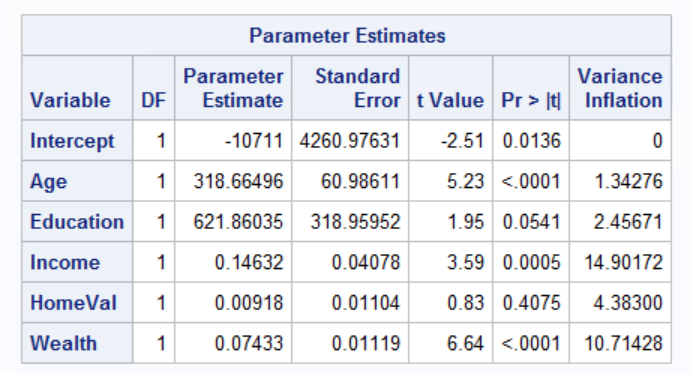
**My findings on the correlation values**: To me this data set looking at the strongly associated values to balance is Income with 0.95168 for a scale of maximum 1 or 95.168% on a scale to 100%. The next strongly associated variable is Wealth at 0.9471 on a scale of 1 or 94.871% on a scale out of 100%.

I would say that HomeVal is a not a strong correlation value that if when looking at scatterplots if those correlation values for a strong correlation is 0.8 which 0.76639 is less than 0.8 meaning to me I cannot call it a strong so to me it is a moderate correlation association.

The last two is Age at 0.56547 or 56.547% is a low correlation and last is Education at 0.55488 or 55.488%. These two have the lowest correlation and not a strong correlation at all.

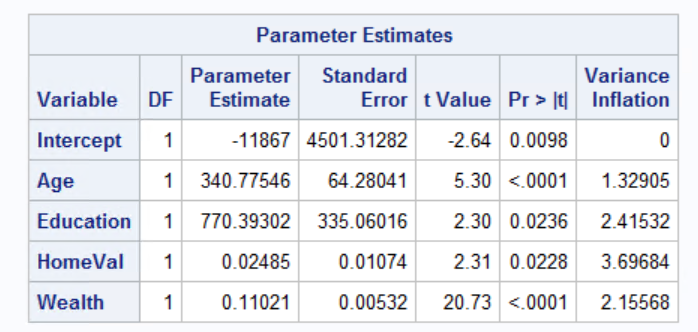
1. Fit a regression model of balance vs the other five variables (model M1). Compute the VIF statistics for each x-variable and analyze whether there is a problem of multicollinearity and take appropriate action. Include the relevant output. Discuss your answer.

**Model 1 VIF** **Statistics:**



**My findings on the VIF**: For the data set I found that running all the x-variables against a VIF >=10 that Income and Wealth are both over 10 against VIF. I want to remove Income first since it has a higher collinearity and re-run the code without Income.

**After dropping Income with the VIF:**

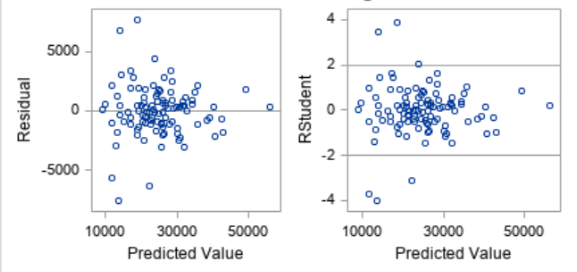


**Updated findings on the VIF**: After dropping Income from the x-variables the rest of the x-variables in the VIF table all went down to under 10 which means there is no more collinearity.

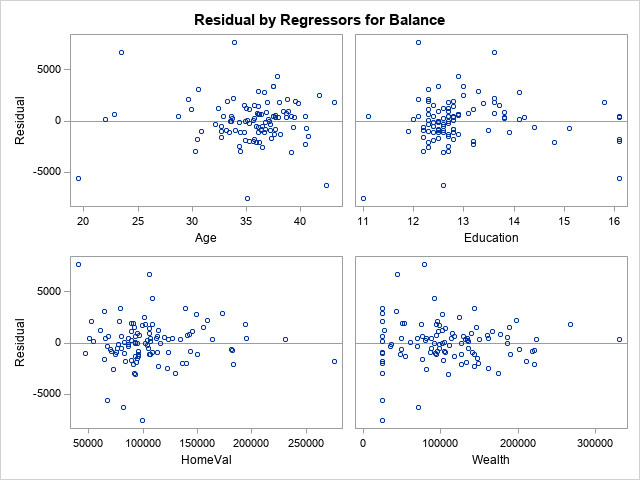
1. Apply your knowledge of regression analysis to define a better model M2. Include the SAS output for both models and answer the following questions :
   1. Analyze the adj-R2 values for both models M1 and M2. Which model has the largest adj-R2 value?

**My findings on adj-R2 values**: For M1 the Adj-R^2 is 0.9441 or 94.41%. For M2 the Adj-R^2 is 0.9373 or 93.73%. The M1 of Model one had a higher Adj-R^2 with 94.41%

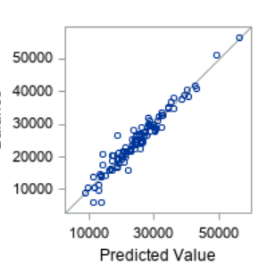
* 1. Create residual plots for M2 (Studentized residuals vs predicted; Studentized residuals vs x-variables; and normal plot of residuals). Analyze the residual plots to check if the regression model assumptions are met by the data. Include the relevant output and discuss your analysis.

**Residuals vs Predicted Plot**: 

**Studentized residuals vs x-variables:**



**Normal Plot of Residuals**:



**My findings on the Residuals vs Predicted Plot:** From the predicted value residual plot I did not see and pattern with the data set because it has a constant starting random flow of data from 0 to where it mostly ends around 30,000. The spread seems constant and random points on the residual plot.

**My findings on the studentized vs x-variables residual plots:**

**Balance and Age:** From what I see of this data setthere can be a funnel shaped pattern on the residual plot because the data does not have many plotted points to begin with and as the Age goes along the data becomes more populated increasing the frequency of data points on the residual plot. Yes there are outliers from the funnel shape I made but reading along the Age x-variable data is scarce then hitting 30-40 the data starts to be plotted and makes for an increasing fashion which makes me think there is a pattern of a funnel.

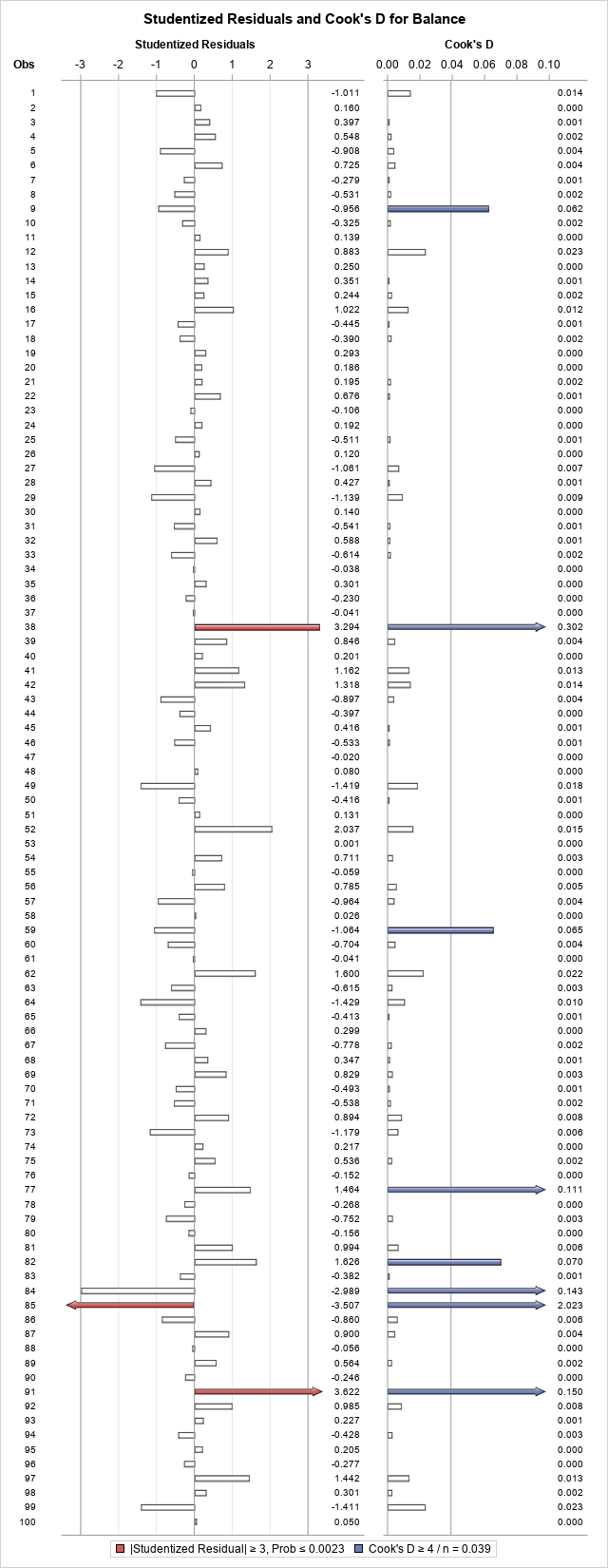
**Balance and Education:** I do not see an obvious pattern to my eyes when I tried a funnel shape or U-shape pattern the spread of the data does not start off small then increase it stays pretty constant throughout the entire Education axis. Why I also believe there is no U-shape pattern because the data does not curve it is random where data starts to dip from where it starts or ended.

**Balance and HomeVal:** From this data I do not see a pattern from this residual plot because data is constant and scattered randomly around the zero line.I do not see a pattern because this data is around the zero-line going across the x-axis HomeVal and points are randomly put with the y-axis a few being above 5000 and some being below -5000 which makes this data random with no pattern.

**Balance and Wealth:** From this data set I do not see an obvious pattern where the data makes a shape. I think the spread of the data is rather constant and random because there is no funnel shape where the data starts off rather small then increases to bigger values. There is no pattern of a U-shape because the data does not curve either way during the x-axis wealth.

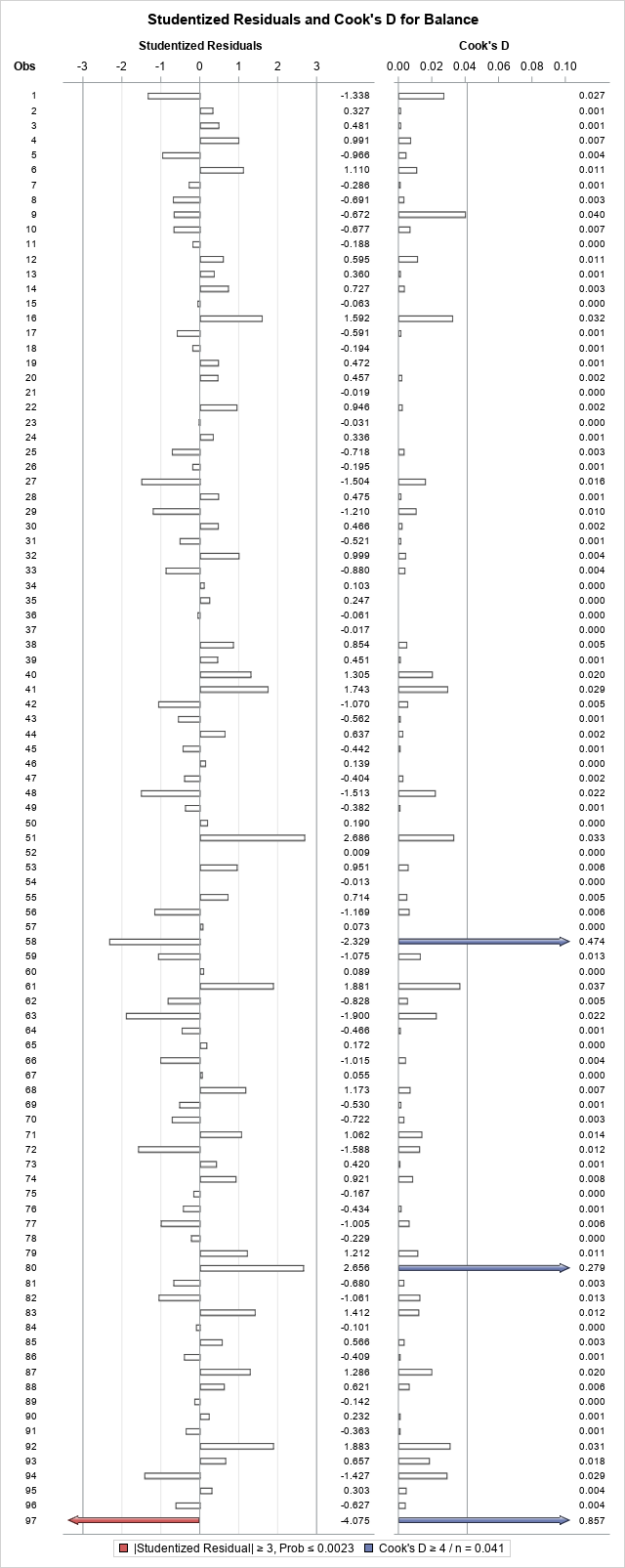
**My findings on the Normal Plot of Residuals:** When I look at this residual plot I did not find anything in this data set that would make me say there is an outlier. The Normal Plot of Residuals the points doe lie close to the line. There is not curve to the residual or a pattern that would make data stray away from the line so no outlier.

* 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.



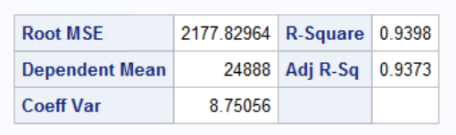
**Influential Points and Outliers in M2**: From running the influential and outliers regression model I found that observations 38 (Positive influential point), 77(Positive influential point), 84(Positive influential point), 85 (Negative outlier and positive influential point), 91 (Both positive outlier and influential point arrows) had arrows in their model so they will need to be removed from the data set and re-ran.

**After removal of observations 38, 77, 84, 85, and 91:**

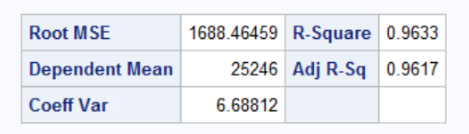


**My findings on outliers and influential points**: I found that after removing these observations that there are now more outliers and influential points that are in this data set. I would say now remove observation 58 (positive influential point), 80 (positive influential point), 97 (Negative outlier and positive influential point). If the Adj-R^2 does not continue to go up either it goes down or not to a value of 3-5% increase then stop and go back to the second model and just use it as proof why I stopped from removing a second time from this data set. After testing and removing and looking back at the regression of the initial data the Ajd-R^2 of the model before influential points or outliers were removed was 0.9373. After I removed the first set of outliers and influential points the Ajd-R^2 was increased to .09617. I tried to remove outliers and influential points again just to see what the future Ajd-R^2 would be if there would be further improvement and it cam back going to 0.9372 meaning that stopping at just the first removal statement is where the regression line with minimal spread is best.

**Adj-R^2 Initial**:



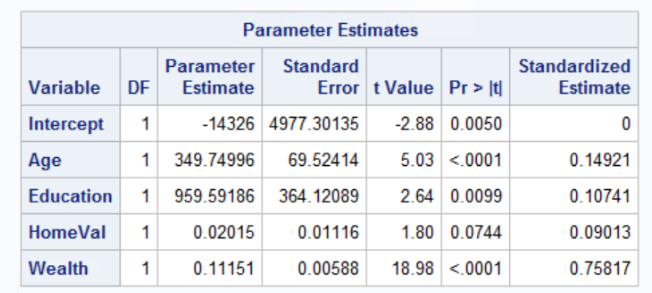
**After removal of 1st set of outliers/influential points**:



**Tried a 2nd set of outliers/influential points:**



* 1. Compute the standardized coefficients for M2 and discuss which predictor has the strongest influence on balance? Include the relevant output.



**My findings on the Standardized Estimate:** From looking at the x-variables to me it tells me that Wealth with 0.75817 or 75.817% has the strongest influence on balance.

Followed by Age at 0.14921 or 14.921% then followed by Education at 0.10741 or 10.741% and finally last is HomeVal at 0.09013 or 9.013%.

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

\*PROC IMPORT;

**PROC** **IMPORT** DATAFILE="Bankingfull.txt" out=banking replace;

DELIMITER='09'x;

GETNAMES=yes;

DATAROW=**2**;

**RUN**;

**PROC** **PRINT** data=banking;

**RUN**;

TITLE"Banking";

**PROC** **PRINT**;

**RUN**;

\*Scatterplot for Balance and Age;

**PROC** **GPLOT**;

PLOT Balance\*(Age);

**RUN**;

\*Scatterplot for Balance and Education;

**PROC** **GPLOT**;

PLOT Balance\*(Education);

**RUN**;

\*Scatterplot for Balance and Income;

**PROC** **GPLOT**;

PLOT Balance\*(Income);

**RUN**;

\*Scatterplot for Balance and HomeVal;

**PROC** **GPLOT**;

PLOT Balance\*(HomeVal);

**RUN**;

\*Scatterplot for Balance and Wealth;

**PROC** **GPLOT**;

PLOT Balance\*(Wealth);

**RUN**;

\*Correlation Values;

**PROC** **CORR**;

VAR Balance Age Education Income HomeVal Wealth;

**RUN**;

\*Regression with VIF;

**PROC** **REG**;

MODEL Balance=Age Education Income HomeVal Wealth/VIF;

**RUN**;

\*Regression with VIF without Income;

**PROC** **REG**;

MODEL Balance=Age Education HomeVal Wealth/VIF;

**RUN**;

\*Regression model to Predict Balance;

**PROC** **REG**;

MODEL Balance=Age Education HomeVal Wealth;

\*Residual Plot;

PLOT student.\*predicted.;

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

PLOT npp.\*student;

**RUN**;

\*Run model with outlier-use original;

**PROC** **REG**;

MODEL Balance=Age Education HomeVal Wealth/influence r;

PLOT student.\*(Age Edcuation HomeVal Wealth predicted.);

PLOT npp.\*student.;

**RUN**;

\*Removal of outliers and influential points;

**data** newbanking;

set banking;

if \_n\_ in (**38**, **77**, **84**, **85**, **91**) then delete;

**RUN**;

\*Regression model to Predict Balance with removal of outliers and influential points;

**PROC** **REG**;

MODEL Balance=Age Education HomeVal Wealth;

**RUN**;

\*New Model without outlier or influential points;

**PROC** **reg** data=newbanking;

MODEL Balance=Age Education HomeVal Wealth/influence r;

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

PLOT npp.\*student.;

**RUN**;

\*Removal of outliers and influential points part 2;

**data** newbankingtwo;

set newbanking;

if \_n\_ in (**58**, **80**, **97**) then delete;

**RUN**;

\*Regression model to Predict Balance with removal of outliers and influential points;

**PROC** **REG**;

MODEL Balance=Age Education HomeVal Wealth;

**RUN**;

\*New Model without outlier or influential points;

**PROC** **reg** data=newbankingtwo;

MODEL Balance=Age Education HomeVal Wealth/influence r;

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

PLOT npp.\*student.;

**RUN**;

\*Regression with standardized coefficients;

**PROC** **REG**;

MODEL Balance=Age Education HomeVal Wealth/stb;

**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;

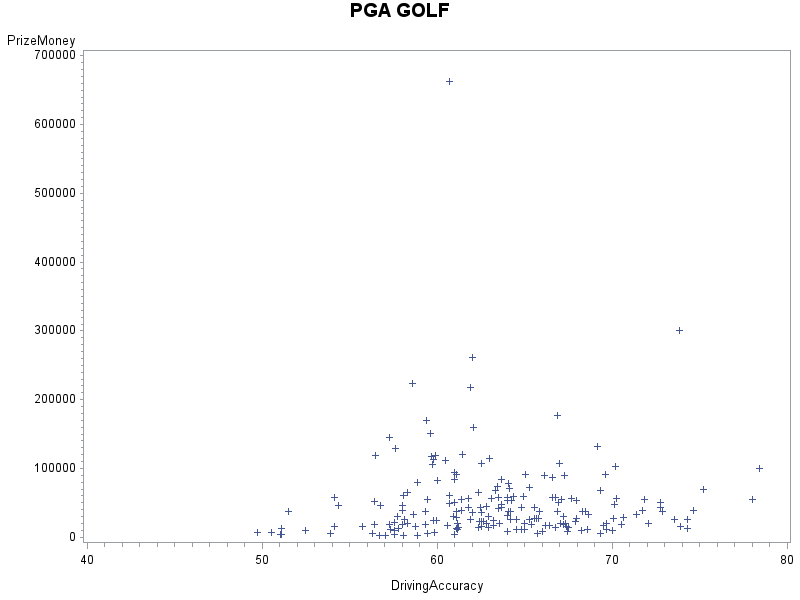
**run**;

Note:

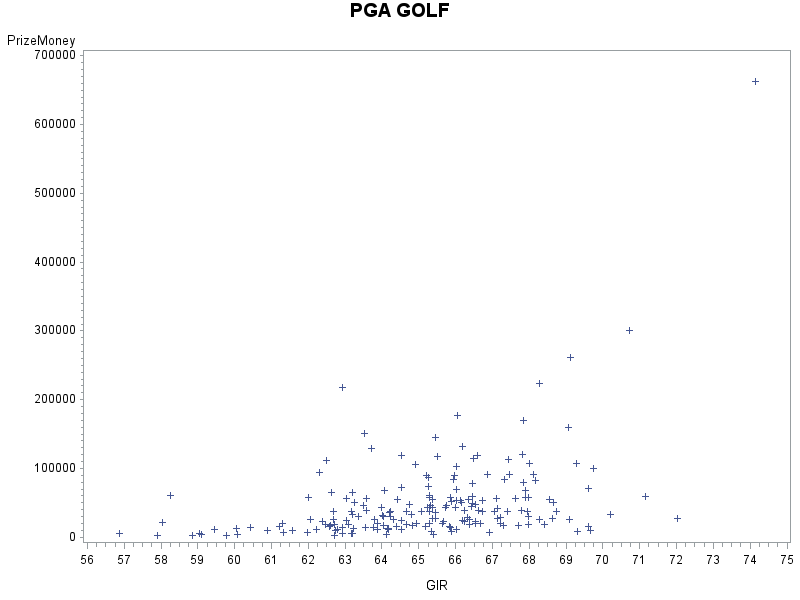
* The data file is in CSV format
* It is delimitered with a comma
* The SAS dataset it is writing into is PGATour. You can change the name if you like.

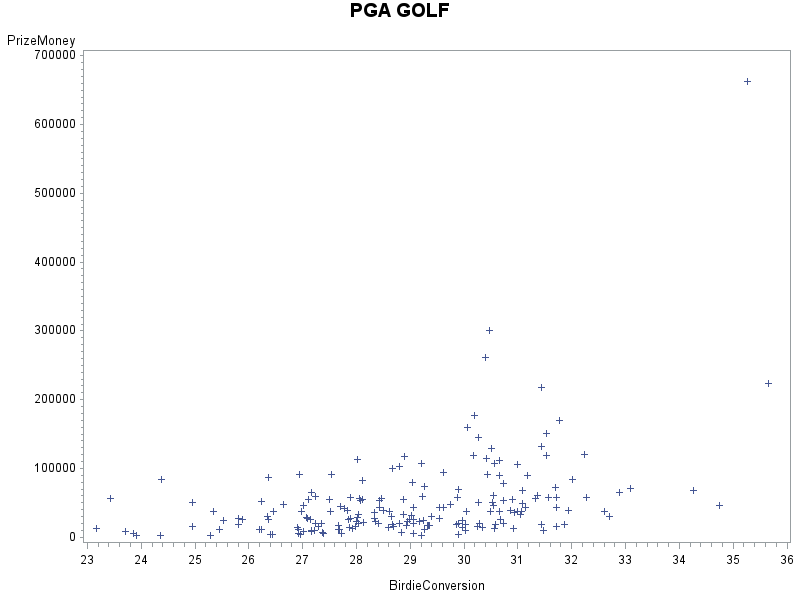
1. Create scatterplots to visualize the associations between PrizeMoney and the other 5 variables. Discuss the patterns displayed by the scatterplot. Also, explain if the associations appear to be linear? (you can create scatterplots or a matrix plot). Include the relevant output.

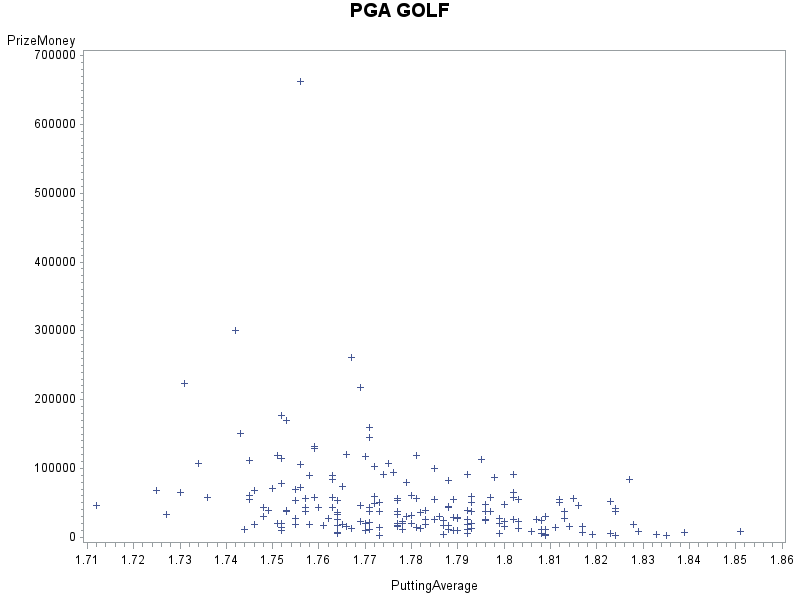
**Scatterplot for PrizeMoney and DrivingAccuracy:**

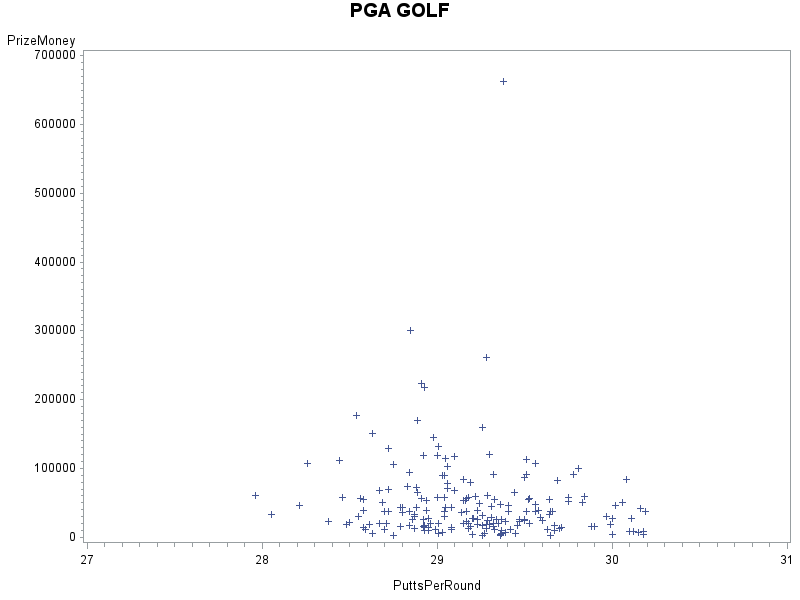


**Scatterplot for PrizeMoney and GIR:**



**Scatterplot for PrizeMoney and BirdieConversion:** 

**Scatterplot for PrizeMoney and PuttingAverage:** 

**Scatterplot for PrizeMoney and PuttsPerRound:** 

**My findings on the scatterplots:**

**PrizeMoney and DrivingAccuracy:** The first graph is a low positive correlation I feel that it is closer to low because of the data when looking at the y-axis from I would say around $100,000 and up that data is pulling away and increasing a spread to this scatterplot which makes me believe it is a low positive correlation. Why I also say logically this is a positive graph because in golf especially professional golf these players should be hitting the green area meaning they hit good far shots more meaning in this graph there should be a positive correlation the more money someone won because of the more percent of shots hit into the fairway.

**PrizeMoney and GIR:** The second graph just like the first graph I believe is another low positive correlation. Why I feel it is not high even though there are a line of points it is everything above $100,000 which makes this spread of data increased and pulling away from the main data values plotted in this scatterplot. Logically again with professional golf if you want to have more winnings and who is at the top because these golfers should be able to hit a greens constantly meaning the more they hit this zone the more earnings they should be getting making this a positive correlation.

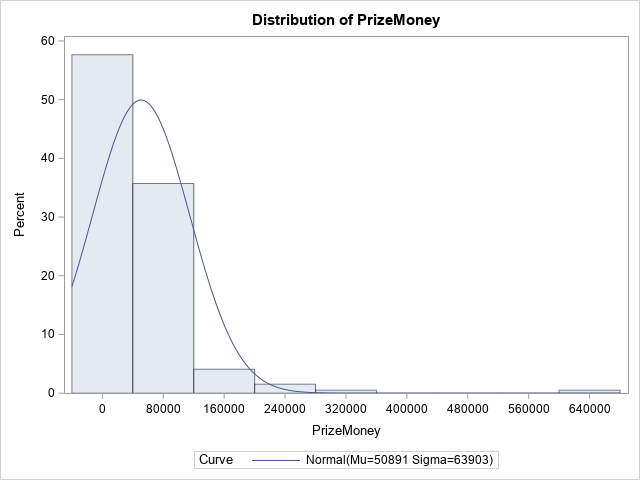
**PrizeMoney and BirdieConversion:** The third graph has no correlation. This graph on the y-axis data is being pulled from the $100,000 mark up making the spread become more vertical. Along with the y-axis the x-axis is being spread out horizontally which makes this graph have this horizonal line going across the entire data but there is never this obvious positive nor negative correlation that I can see.

**PrizeMoney and PuttingAverage:** The fourth graph is a low negative correlation. Here is why though I am not a golf expert or play it at all from what I see and searched that when golfers hit into this green zone they want to convert and should convert at a higher percent of their putts when in this zone. From the data I see a majority of the plotted points in 1.73 to 1.79 with the few expectations that people are able to hit at 1.81-1.84 but very far less in comparison to the first 1.73-1.79 making this a low negative correlation because the data as well is still being spread out on the y-axis when hitting $150,000-$300,000.

**PrizeMoney and PutsPerRound**: The fifth graph to me is a low negative correlation. In gold especially at a professional level if a golfer has a chance to be putting which means closer shots to the hole they should be converting those close ranged putts. What this data is showing to me that the majority of these golfers do it in 28..5-29.5 putts per round. The majority of the data is there and starts to become less plotted when heading towards 30 and even 30+ putts per round. I see the better the golfer is the more they are able to covert on those close ranged putts.

1. Analyze distribution of PrizeMoney, and discuss if the distribution is symmetric or skewed. Include the relevant output.

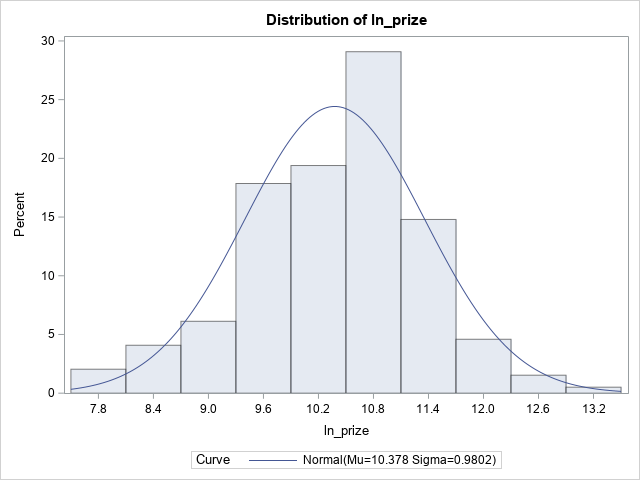
**Histogram to find distribution of PrizeMoney:**



**My findings on the distribution of PrizeMoney:** After making the histogram I found that this distribution of the PrizeMoney is positively skewed or a right skewed histogram. Since all the PrizeMoney starts at $0 which also happens to be the peak of this distribution then starts to trend downwards to $80,000 eventually hitting $640,000 skewing the entire histogram pulling the plotted line all the way right of the main data.

1. 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.

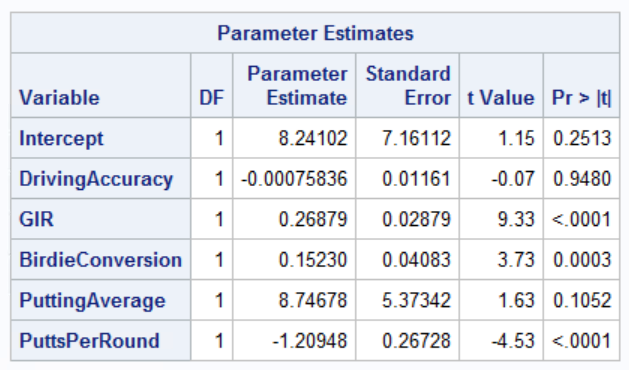
**Histogram of ln\_prize:**



**My findings of ln\_prize**: I find that the distribution of the histogram of ln\_prize is rather normal with one peak unimodal flow of data which rises from 0, peaks at 10.8, and goes down to 13.2. There is no visible outliers on this histogram.

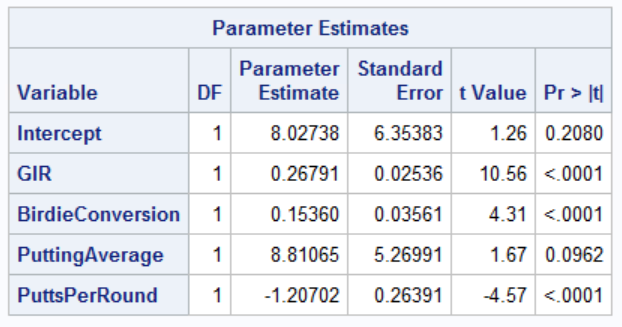
1. Fit a regression model of ln\_Prize using the remaining predictors in your dataset. Apply your knowledge of regression analysis to define a valid model to predict ln\_Prize. Include the outputs for all the questions below before you analyze them.
   * 1. If necessary remove the non-significant variables. 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.

**Regression model of ln\_prize**:



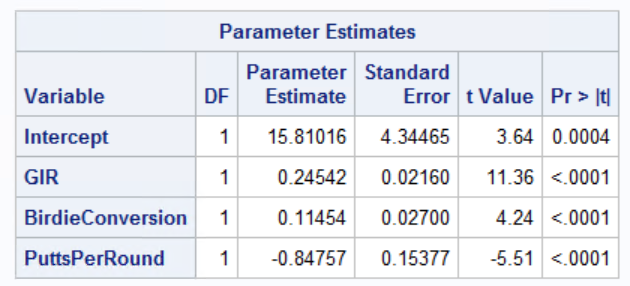
**My findings on this regression model**: I find that looking at the alpha test=0.05 that DrivingAccuracy is 0.9480 which is greater than 0.05 meaning this variable is not significant and this model needs to remove it and re-run this regression model.

**Updated Regression Model without DrivingAccuracy**:



**My findings on this regression model**: I find that looking at the alpha test=0.05 that PuttingAverage is 0.0962 which is greater than 0.05 meaning this variable is not significant and this model needs to remove it and re-run this regression model.

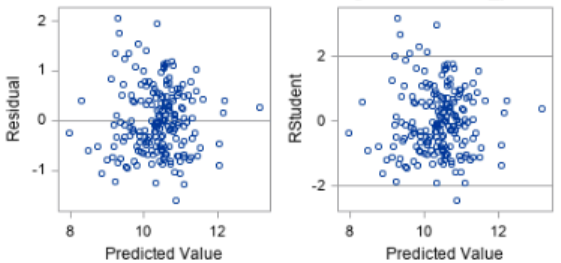
**Updated Regression Model without PuttingAverage**:

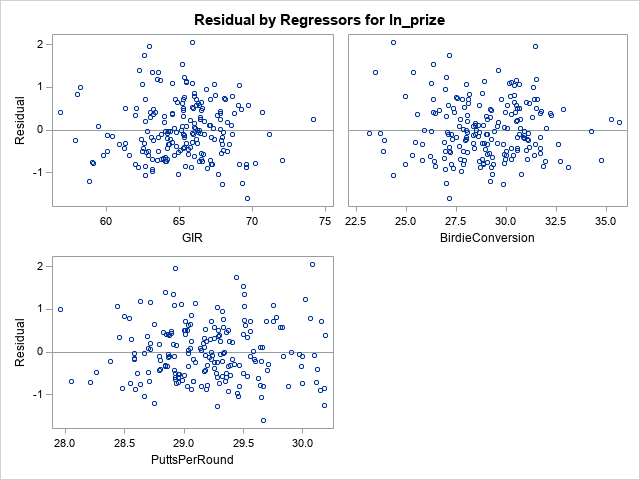


**My findings on the regression model**: When looking at the alpha test=0.05 there all of the x-variables GIR, BirdieConversion, and PuttsPerRound are all <.0001 which is less than 0.05 meaning all three x-variables are significant.

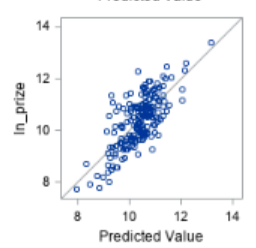
* + 1. Analyze residual plots to check if the regression model is valid for your data. Discuss your analysis.

**Residuals vs Predicted Plot:**



**Studentized vs x-variables Plot:** 

**Normal Plot of residuals:**



**My findings on the Residuals vs Predicted Plot:** I think that there is not pattern since the data is constant spread across the line and has random points making no obvious pattern to how it is plotted. I tried a funnel and U-shaped pattern but neither work funnel is too wide shaped and U-shaped pattern is not very U-shaped.

**My findings on the studentized vs x-variables residual plots:**

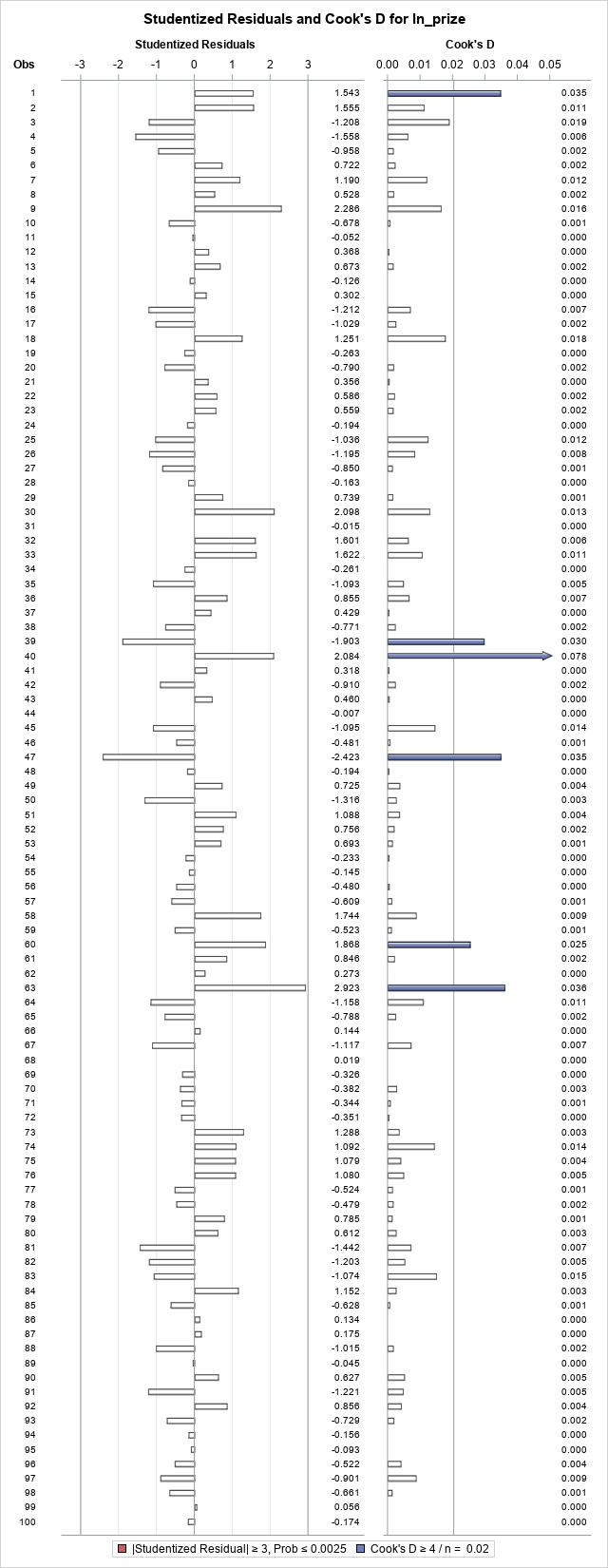
**ln\_prize and GIR:** There is no pattern to the residual plot all the points are random and the spread is constant throughout the residual plot. I do not see a funnel shaped curve because the data throughout the entire plot is random and no way there is an increase or even a decrease pattern. There is no U shape because the graph does not curve in one way or another.

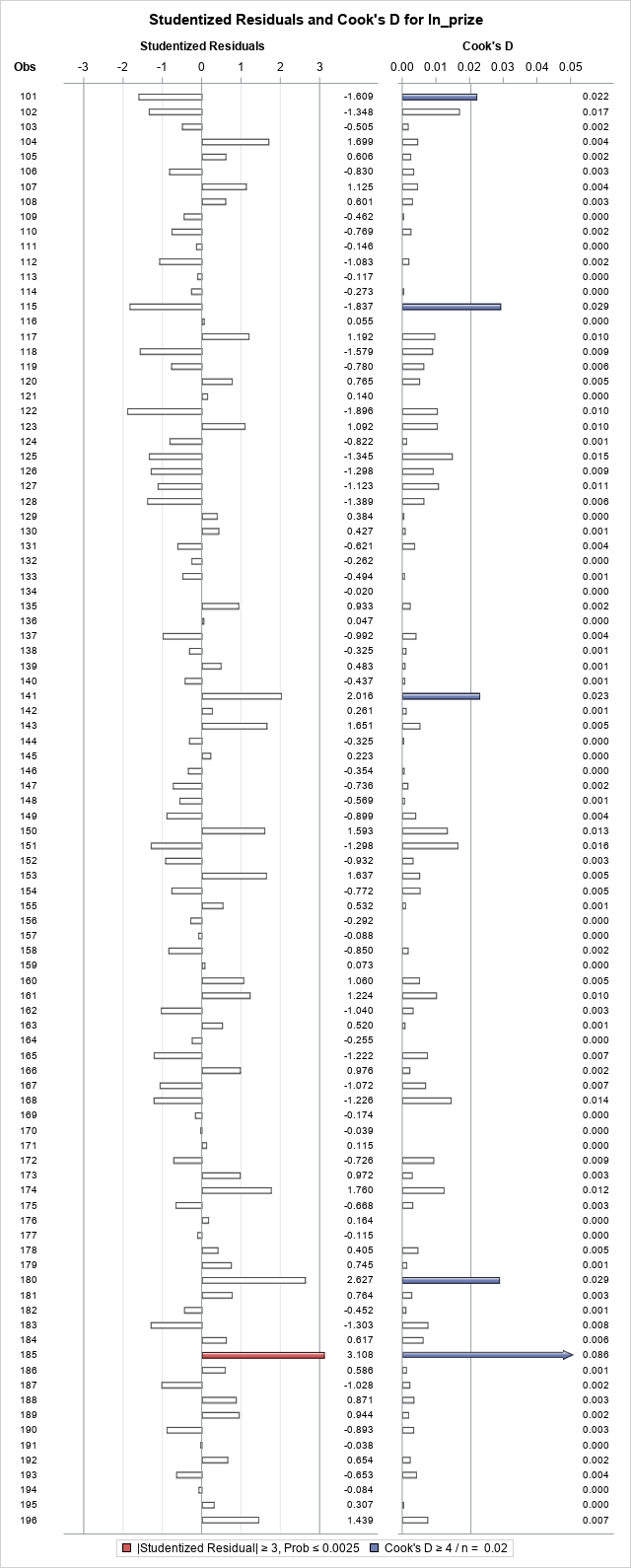
**ln\_prize and BirdieConversion:** This residual plot does not have a pattern and is a spread constant where there is random elements of this graph. There is not funnel shape because the plot goes at random points spread out and no U shape because the plot does not go inward or outward to make a U shape plot.

**ln\_prize and PuttsPerRound:** This residual plot does not have a pattern and the spread of the plot is constant where I feel there is not pattern like a funnel or U-shaped pattern. I believe the data is constant spread and points are random. I tried to make a visual funnel or U shaped but this would be too wide for a funnel making a very wide funnel if so and U shape does not work.

**My findings on the Normal Plot of Residuals:** I think that this normal plot of residualshas a good linear close spread of data that would tell me there is no outlier or influential point in this data set. All the points on the Normal Plot lie close to the line and no curve or pattern to the data.

* + 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.



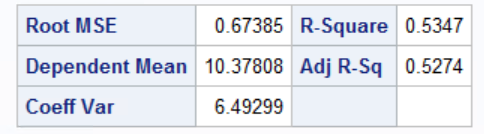


**My findings on Outliers/Influential points on PGATour2006.csv**:

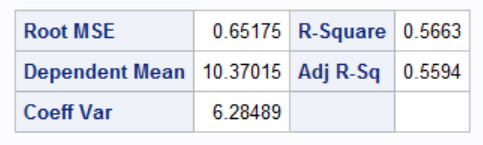
After running for outliers/influential points on ln\_prize I found that 40 (positive influential point) and 185 (positive influential point) were to be removed.

Also seeing the Adj-R^2 going up means that the data should have been removed. After re-running the influential points and outliers charts there was no indication of anymore arrows with either red or blue for ln\_prize on any of the observations so I believe there is no more influential or outliers in this data set.

**Initial Ajd-R^2:**



**After first removal Ajd-R^2:**



* + 1. Write down the final model equation. Discuss why this is the best model. Include all relevant statistics/values to substantiate your answer.

**The model equation is (rounded up or down from .5):**

ln\_prize=14.73 + 0.24 GIR + 0.14 BirdieConversion -0.83 PuttsPerRound

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?

**GIR**: It is a positively associated with ln\_prize. Model shows that assuming all other variables are constant, for every additional 1% GIR, ln\_prize increases by 27.12%%.

logOpen=exp(0.24)-1\*100=27.12%

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"PGA GOLF";

**PROC** **PRINT**;

**RUN**;

\*Scatterplot for PrizeMoney and DrivingAccuracy;

**PROC** **GPLOT**;

PLOT PrizeMoney\*(DrivingAccuracy);

**RUN**;

\*Scatterplot for PrizeMoney and GIR;

**PROC** **GPLOT**;

PLOT PrizeMoney\*(GIR);

**RUN**;

\*Scatterplot for PrizeMoney and BirdieConversion;

**PROC** **GPLOT**;

PLOT PrizeMoney\*(BirdieConversion);

**RUN**;

\*Scatterplot for PrizeMoney and PuttingAverage;

**PROC** **GPLOT**;

PLOT PrizeMoney\*(PuttingAverage);

**RUN**;

\*Scatterplot for PrizeMoney and PuttsPerRound;

**PROC** **GPLOT**;

PLOT PrizeMoney\*(PuttsPerRound);

**RUN**;

\*Histogram to find distribution;

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

VAR PrizeMoney;

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

**RUN**;

\*Create log prize variable in a data step;

**data** PGATour;

\*Set command copeis orginal PGATour dataset;

set PGATour;

\*Log transformation of PrizeMoney;

ln\_prize=log(PrizeMoney);

**PROC** **PRINT**;

**RUN**;

\*Histogram of ln\_prize;

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

VAR ln\_prize;

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

**RUN**;

\*Regression Model of ln\_prize;

**PROC** **REG**;

MODEL ln\_prize=DrivingAccuracy GIR BirdieConversion PuttingAverage PuttsPerRound;

**RUN**;

\*Regression Model of ln\_prize without DrivingAccuracy;

**PROC** **REG**;

MODEL ln\_prize=GIR BirdieConversion PuttingAverage PuttsPerRound;

**RUN**;

\*Regression Model of ln\_prize without PuttingAverage;

**PROC** **REG**;

MODEL ln\_prize=GIR BirdieConversion PuttsPerRound;

**RUN**;

\*Residual Plot of ln\_prize;

**PROC** **REG**;

MODEL ln\_prize=GIR BirdieConversion PuttsPerRound;

\*Residual Plot;

PLOT student.\*predicted.;

PLOT student.\*(GIR BirdieConversion PuttsPerRound);

PLOT npp.\*student;

**RUN**;

\*Run model with outlier-use original;

**PROC** **REG**;

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

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

plot npp.\*student.;

**RUN**;

\*Removal of outliers and influential points;

**data** newPGATour;

set PGATour;

if \_n\_ in (**40**, **185**) then delete;

**RUN**;

\*Regression Model of ln\_prize without PuttingAverage;

**PROC** **REG**;

MODEL ln\_prize=GIR BirdieConversion PuttsPerRound;

**RUN**;

\*New model without outlier/influential points;

**PROC** **REG**;

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

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

plot npp.\*student.;

**RUN**;