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

**Assignment-4** | **Total Points: 26 (CSC 423 and CSC 324)**

**Due Date: 02/13/2018 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.***

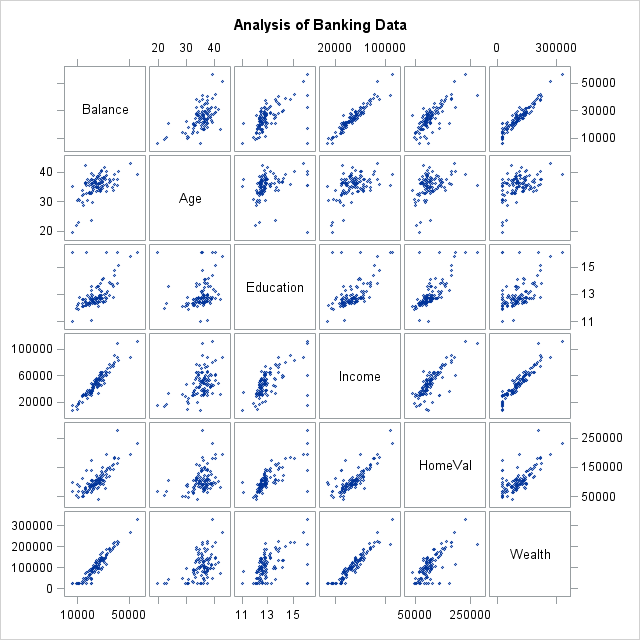
**PROBLEM 1 [16 pts] – to be answered by everyone**

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 $

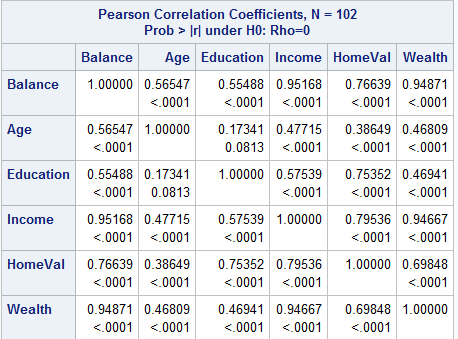
The goal of this exercise is to define a regression model to predict the average bank balance as a function of the other variables.

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)



The association between balance and income, homeval, and wealth appears to be highly positively linear. The association between balance and education appears to be slightly linear. The association between balance and age does not appear to be linear.

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.



**Balance and Income have the strongest correlation, followed by balance and wealth. Balance and homeval have a slight correlation. Balance and age and balance and education have poor correlation.**

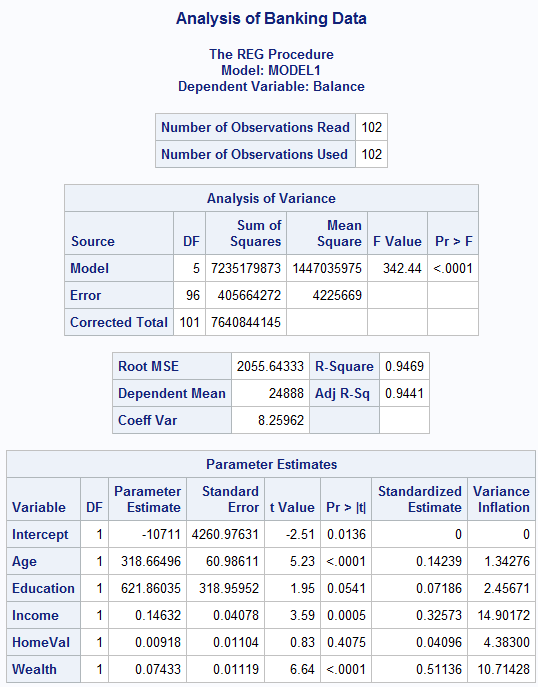
**Also, wealth and income are highly correlated.**

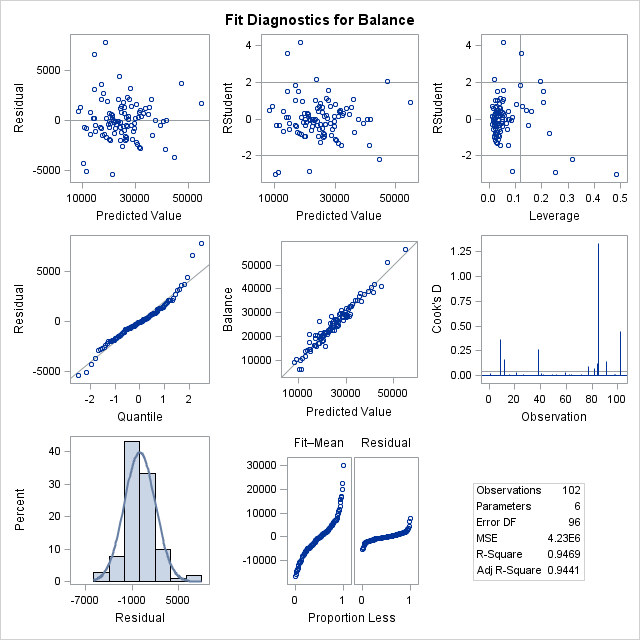
1. Fit a regression model of balance vs the other five variables (model M1). Compute the VIF statistics for each x-variable and analyze whether there is a problem of multicollinearity and take appropriate action. Include the relevant output. Discuss your answer.

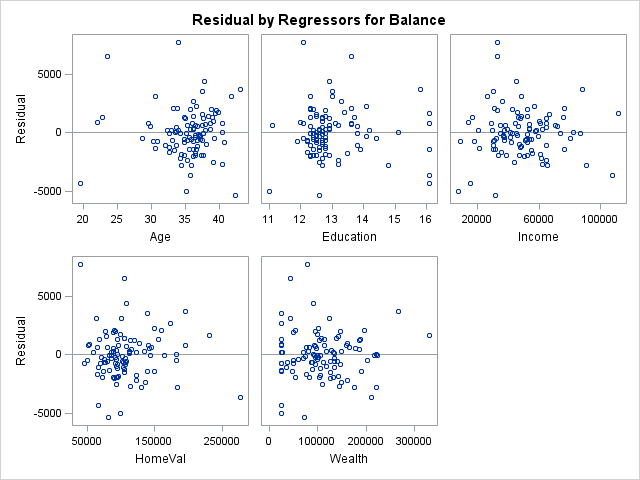
**As we saw with the correlation, income and wealth have a VIF > 10. In this case, I would ask the client if they want me to keep both variables in the model. If one is to be removed, I would remove income because it has a lower standardized estimate than wealth.**

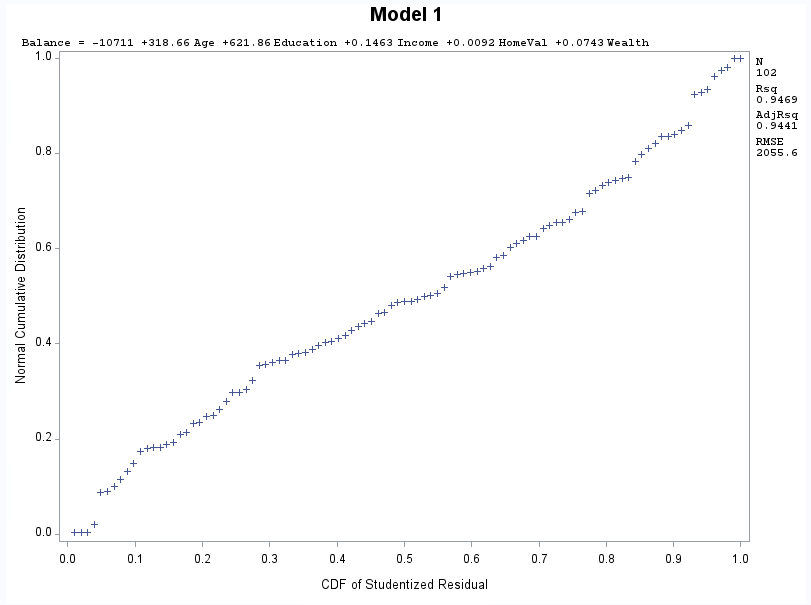
**I would also remove homeVal because this variable has a p-values > 0.05, meaning it does not have a large influence on the balance. Overall the model is a good fit according to the F Value.**

Model 1: Outlined in RED





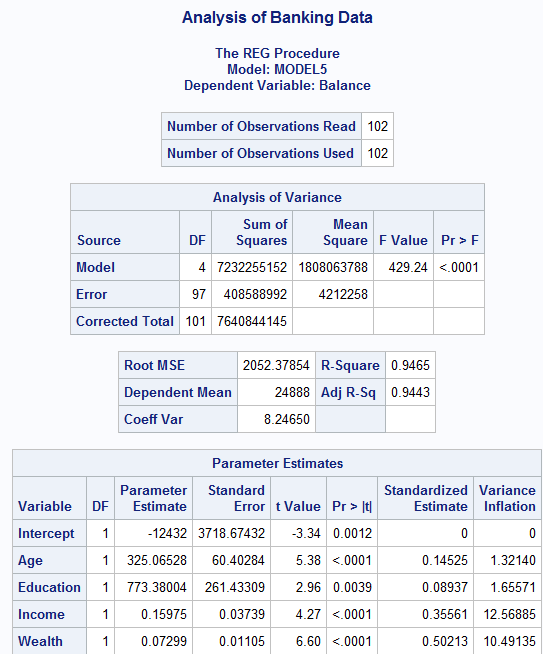


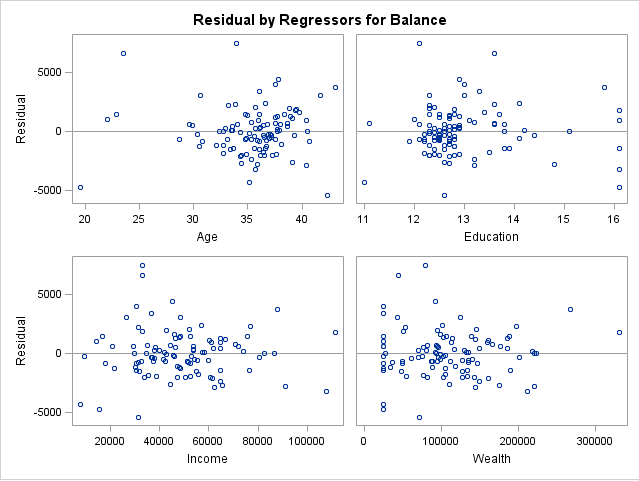
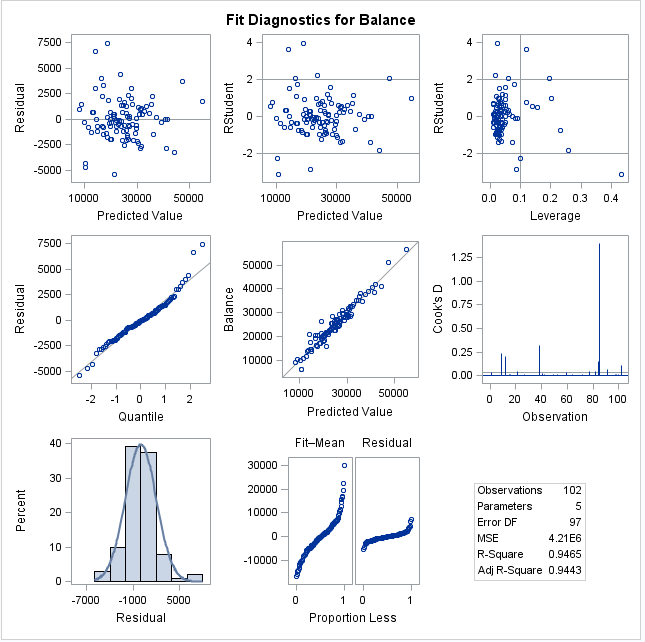


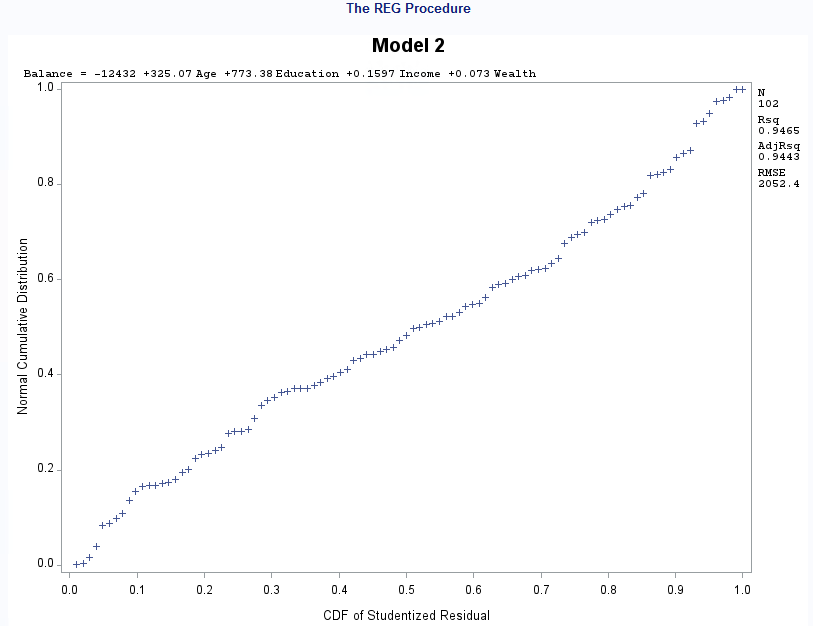
1. Apply your knowledge of regression analysis to define a better model M2. Include the SAS output for both models and answer the following questions :

Model 1: Outlined in RED

Model 2: Outlined in Blue







* 1. Analyze the adj-R2 values for both models M1 and M2. Which model has the largest adj-R2 value?

**The adj-R2 value for M2 is greater than M1. This is due to removal of the homeval variable.**

* 1. Create residual plots (standardized residuals vs predicted; standardized 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.

**For both M1 and M2:**

**Constant variance, independence and linearity are violated.**

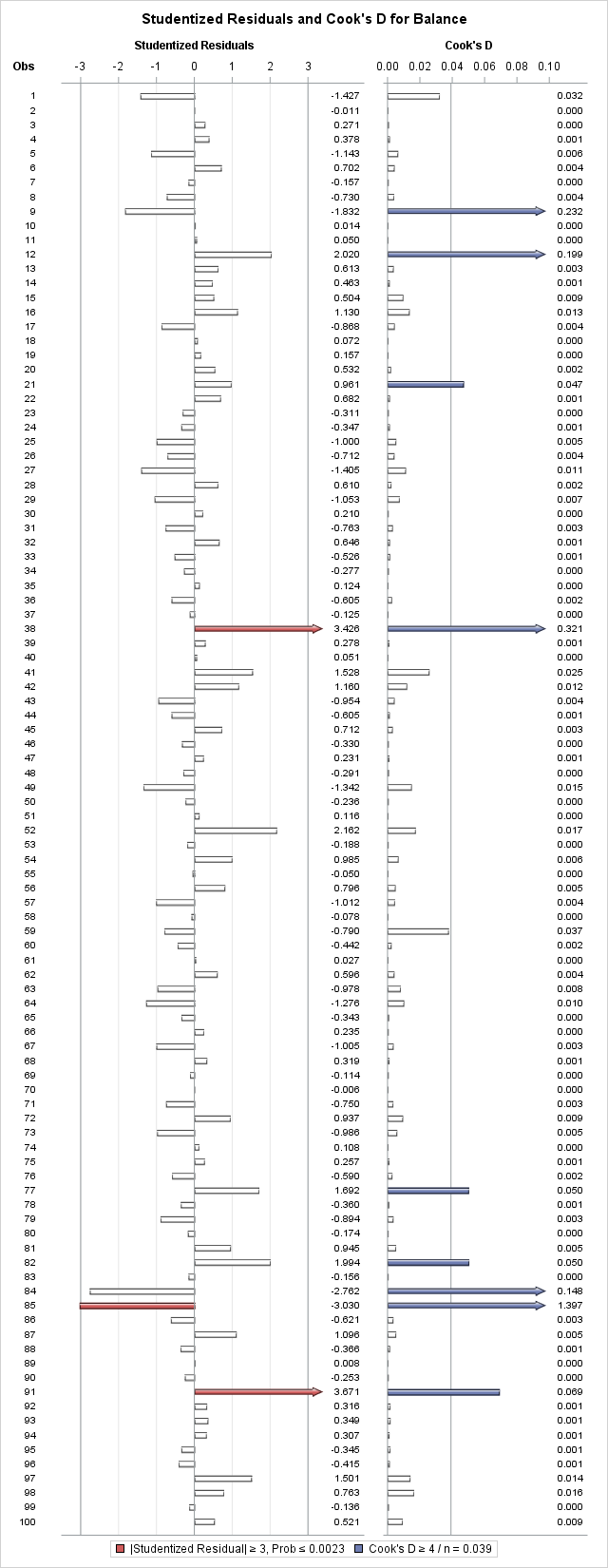
**Both appear to be fairly normal.**

* 1. Analyze if there are any outliers and/or influential points for your 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.

**Yes, there are some outliers and influential points.**

**Data point 38, 85, 91 are outliers**

**Data points 9, 12, 21, 38, 77, 82, 84, 85, 91, 102 are influential points.**



**I would remove points 85 and 91 because they are both outliers and influential points.**

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

**Wealth has the strongest influence on the balance followed by income.**

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

Title "Analysis of Banking Data";

**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';

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

**run**;

plot npp.\*student.;

**run**;

**Problem 2 [10 pts] – to be answered by everyone**

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=myd replace;

delimiter=',';

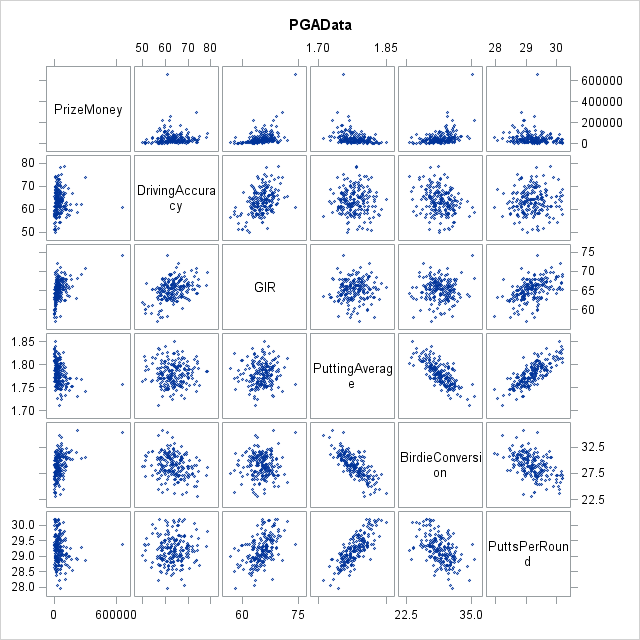
getnames=yes;

**run**;

Note:

* The data file is in CSV format
* It is delimitered with a comma
* **The SAS dataset it is writing into is myd. 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.



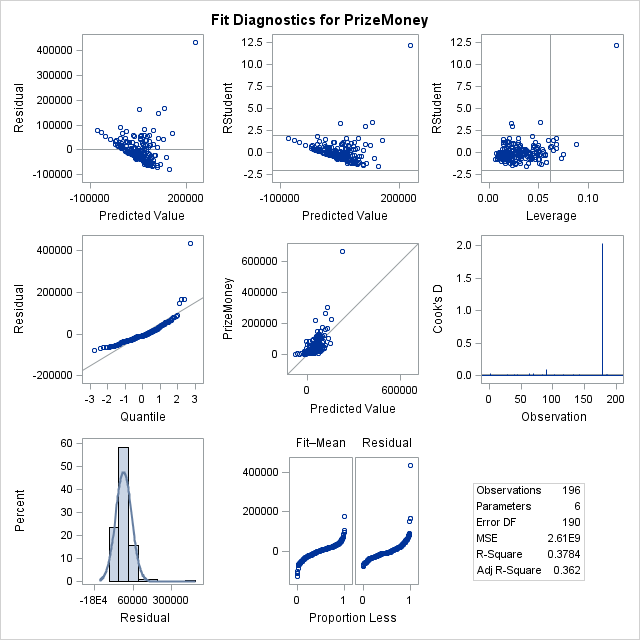
**PuttingAverage and BirdieConversion appear to have a negative linear relationship.**

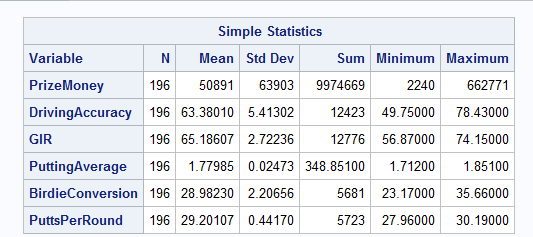
**PuttingAverage and PuttsperRound appear to have a positive linear relationship.**

**BirdieConversion and PuttsPerRound have a negative linear relationship.**

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

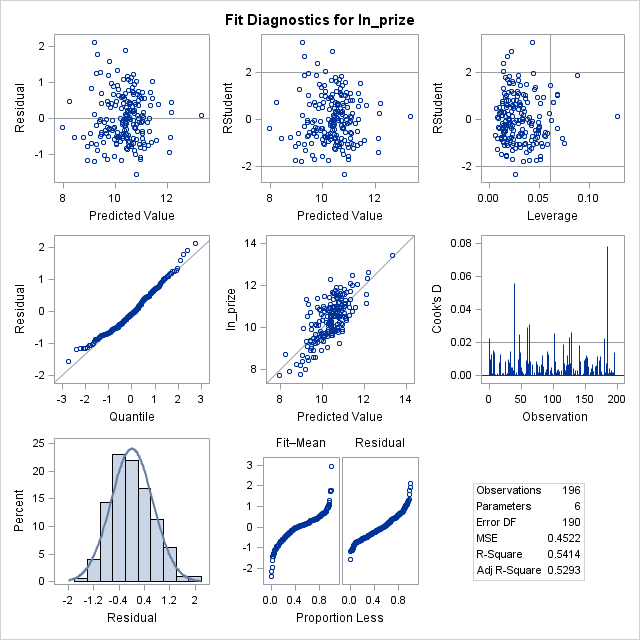
**The distribution of prizemoney is skewed right.**



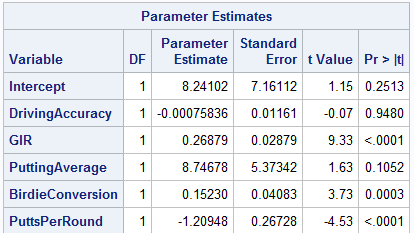


1. Apply a log transformation to PrizeMoney and compute the new variable ln\_Prize=log(PrizeMoney). Analyze distribution of ln\_Prize, and discuss if the distribution is symmetric or skewed. Include the relevant output.

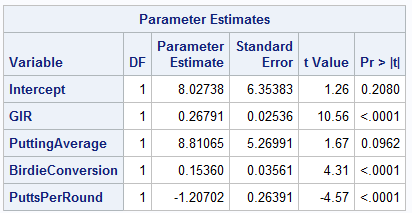
**The distribution of ln\_prize appears to be normally distributed.**



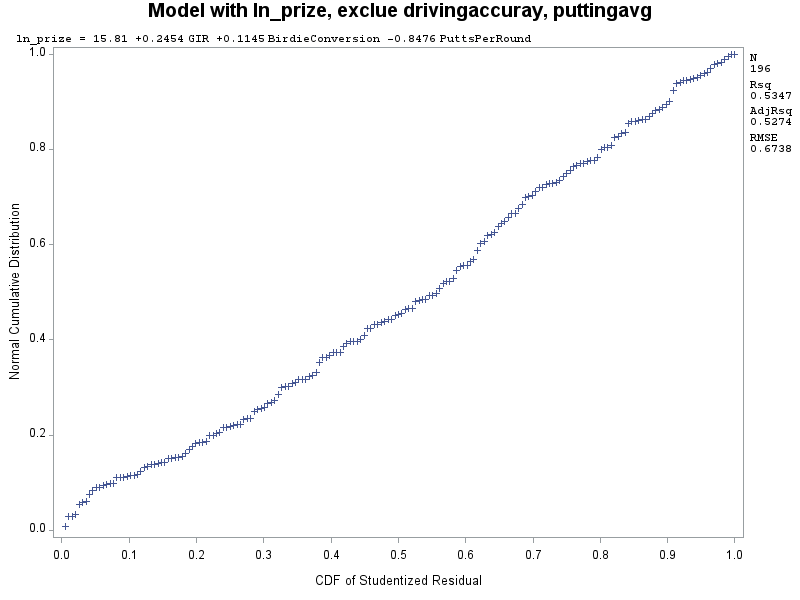
1. Fit a regression model of ln\_Prize using the remaining predictors in your dataset. Apply your knowledge of regression analysis to define a valid model to predict ln\_Prize. Include the outputs for all the questions below before you analyze them.
   * If necessary remove the non-significant variables. 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.
     1. **Start with removing Driving Accuracy, with p-value = 0.9480**

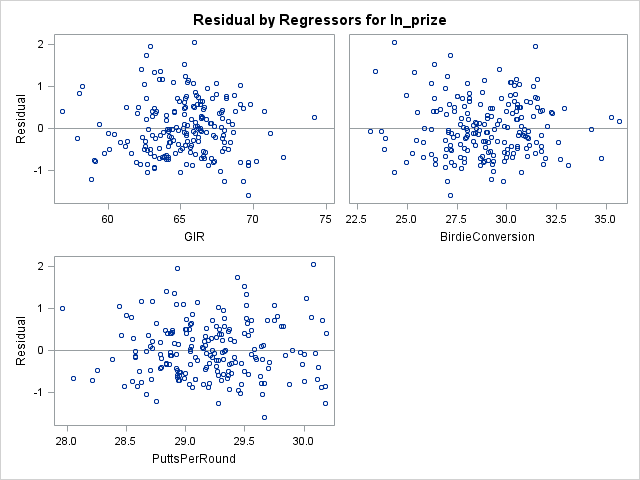


* + 1. **Second, remove Putting Average, with p-value = 0.0962**



* + Analyze residual plots to check if the regression model is valid for your data. Discuss your analysis.
    1. **Residual plots have constant variance, normality, and linearity. It looks as if they have some type of pattern, which would violate independence.**





* + 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.

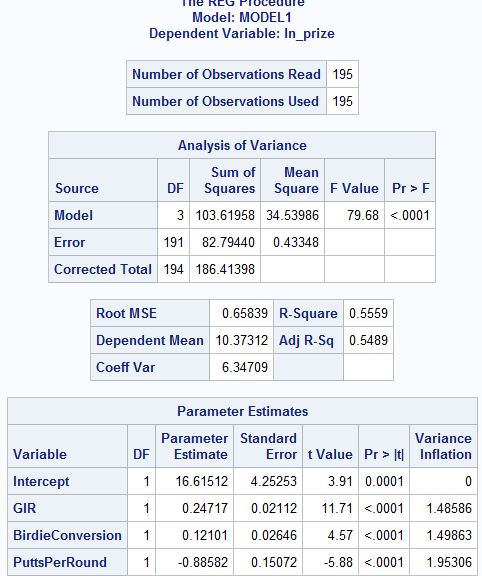
**Value 185 is an outlier**

**Influential points: 1, 40, 41, 47, 60, 63, 101, 115, 141, 180, 185**

**I will remove point 185 because it is both an outlier and influential point.**

**Removing point 185 resulted in a higher F-value, lower Root MSE, and high R2 and Adj R2.**

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



**Ln\_prize = 16.615 + 0.247GIR + 0.121BirdieConversion – 0.885PuttsPerRound**

**This is the best model because all variables are significant, the f-value is high, there is not multicollinearity, and it has the best AdjR2.**

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?

**1% increase in GIR leads to an increase of prize money by 28%**

**You need to make sure you take the exponent of GIR because the original equation is in ln\_prize money.**

**Prize money = Exp((0.247) -1 )\*100**

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

Title 'PGAData';

\*import data from file;

**proc** **import** datafile="S:\Homeworks\pgatour2006.csv" out=myd replace;

delimiter=',';

getnames=yes;

**run**;

**proc** **print**;

**run**;

**proc** **sgscatter**;

matrix prizemoney drivingaccuracy gir puttingaverage birdieconversion puttsperround;

**run**;

**proc** **corr**;

var prizemoney drivingaccuracy gir puttingaverage birdieconversion puttsperround;

**run**;

Title 'Model 1';

**proc** **reg**;

model prizemoney = drivingaccuracy gir puttingaverage birdieconversion puttsperround;

**run**;

plot npp.\*student.;

**run**;

**data** myd;

set myd;

ln\_prize = log(PrizeMoney);

**run**;

**proc** **print**;

**run**;

Title 'Model with ln\_prize';

**proc** **reg** data=myd;

model ln\_prize = drivingaccuracy gir puttingaverage birdieconversion puttsperround;

plot student.\*predicted;

plot student.\*(drivingaccuracy gir puttingaverage birdieconversion puttsperround);

**run**;

Title 'Model with ln\_prize, exclue drivingaccuray';

**proc** **reg** data=myd;

model ln\_prize = gir puttingaverage birdieconversion puttsperround;

plot student.\*predicted;

plot student.\*( gir puttingaverage birdieconversion puttsperround);

**run**;

Title 'Model with ln\_prize, exclue drivingaccuray, puttingavg';

**proc** **reg** data=myd;

model ln\_prize = gir birdieconversion puttsperround;

plot student.\*predicted;

plot student.\*( gir birdieconversion puttsperround);

plot npp.\*student.;

**run**;

Title 'Model ln\_prize, exclue drivingaccuray, puttingavg, outliers';

**proc** **reg** data=myd;

model ln\_prize = gir birdieconversion puttsperround/vif influence r;

**run**;

\*removing point 185;

title 'Proc print after removing point 185';

**data** mydnew;

set myd;

\*remove point;

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

**run**;

**proc** **print**;

**run**;

Title 'Model w/o outlier 185';

**proc** **reg** data=mydnew;

model ln\_prize = gir birdieconversion puttsperround/vif influence r;

plot student.\*predicted.;

plot npp.\*student.;

**run**;