**Jon Doretti**

**DATA ANALYSIS AND REGRESSION**

**Assignment-2** | **Total Points: 35 pts for DSC 323; 40 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.***

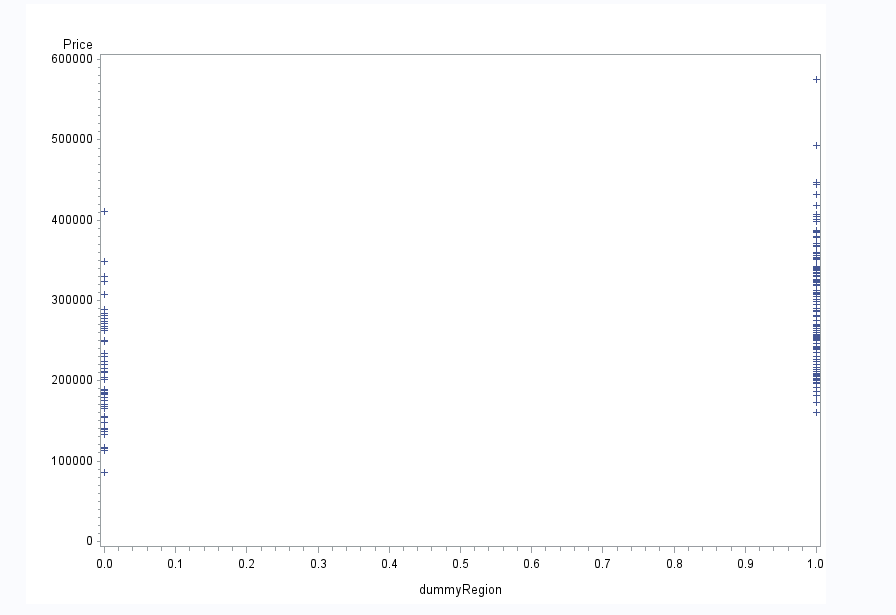
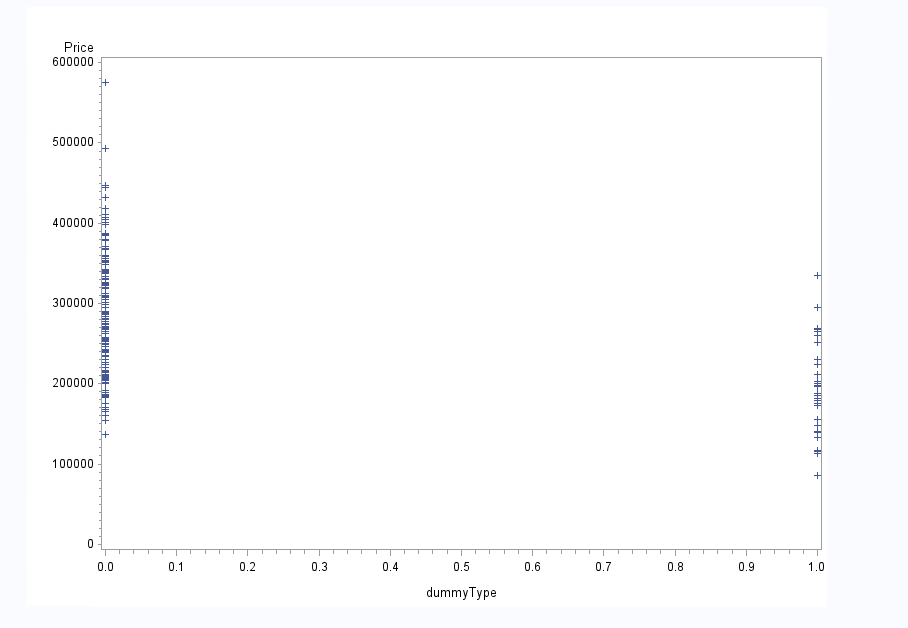
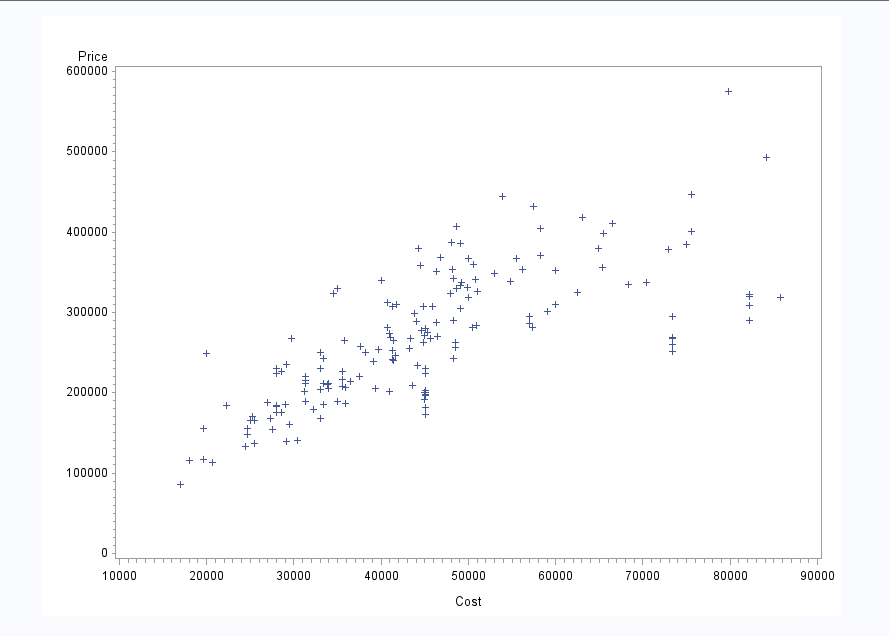
**Problem 1 [15pts] – to be answered by everyone**

A national homebuilder builds single-family homes and condominium style townhouses. The file housesales.txt provides information on the selling price (PRICE), lot cost (COST), type of home (TYPE) (R=row homes home or C=condominium) and region of the country (REGION) (MW=Midwest, S=south) for closings during one month.

a) Define the dummy variables for region and type (write them down here), and create them in SAS. Include your code segment for recoding the variables here.

**Two variables - dummyRegion and DummyType were created. When dummyRegion equals one, it represents the midwest and zero represents the south. When dummyType equals one it equals Row homes home and zero equals Condominium.**

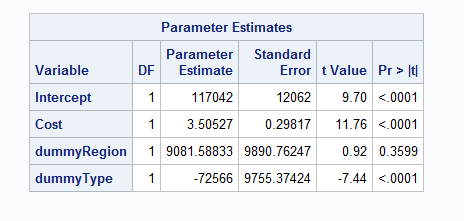
b) Analyze the association between selling price and each individual attribute (cost, type and region) using appropriate statistics and graphs. Discuss your findings. Include the relevant output.



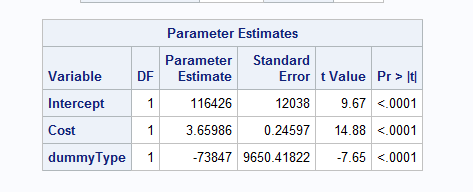
**In the first image there is a low positive correlation between selling price and cost. This is because the range on the y-axis, price, is scattered. However, the range on the x-axi, cost, has a positive correlation of data.**

**The second and third pictures show no linear relation as the data points for the dummy variables are either one or zero. There is no range for them to follow.**

c) Fit an adequate regression model for sales price as a function of lot cost, region of country, and type of home. Remove the terms that are not significant. The final model should only contain variables that are significantly associated with sale price. Write down the model equation. Include the relevant output.



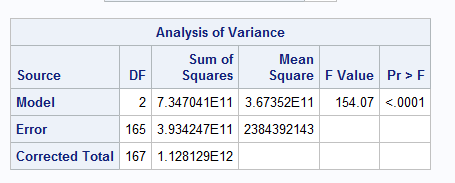
**dummyRegion is not significant because p- value is greater than .05.**



**The above image includes only significant values as all p-values are less than .0001. A variable is considered significant if the p-value is less than .05.**

**Price = 116426 + 3.66 (Cost) 73847 (dummyType)**

d) Conduct a global F-test for overall model adequacy. Write down the test hypotheses and test statistic and discuss conclusions. Include the relevant output.



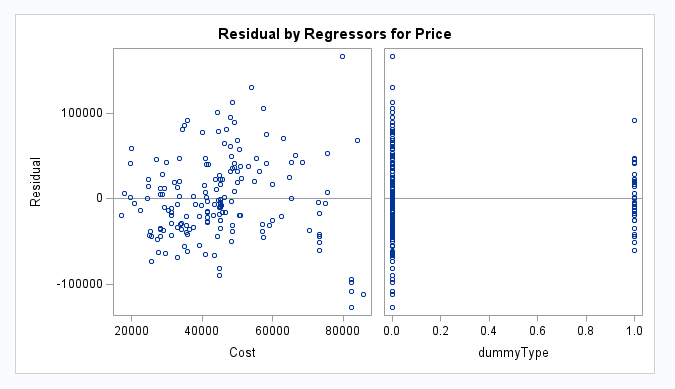
**Null hypothesis - all x variables have no association with any y-variables**

**Alt-hypothesis - At least one x-variable has a significant effect on any y-variable**

3**.67352E11/2384392143 = 154.06**

**P-value = <.0001**

**Here the alt-hypothesis is accepted as the p-value of <.0001 is less than the alpha test value of .05. In other words, the null hypothesis is rejected and at least one x-variable has a significant effect on any y-variable.**

e) Analyze model residuals to check if assumptions on data are satisfied. Discuss your findings. Include the relevant output.

**There is no correlation on the residual plot for cost. With data points scattered throughout the plot you can not draw a funnel or any pattern. Resulting in a constant spread of data. Again, the dummyType residual plot shows no relation at all as there are only two types of dummyType. Resulting in either a 1 or 0.**

f) Discuss what the regression model indicates for the relationship between price and home type (i.e. interpret the coefficient values).

**Price and home type are positively associated. This can be seen with all home types resulting in zero; with each zero, price increases.**

g) Use the regression analysis to determine whether mean sale prices are different for the two regions? Explain.

**There are two different mean sale prices. This is because condominium sales are usually in higher density urban areas. This will result in higher prices - take downtown Chicago for example. While with Row homes home, I am assuming these are prebuilt homes built in scale resulting in a lower cost due to manufacturing in bulk. These two differences will change the mean sale prices. Along with selling in different areas.**

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

PROC IMPORT DATAFILE="C:\Users\JDORETTI\Downloads\HouseSales.txt" out=Sales\_import replace;

DELIMITER='09'x;

GETNAMES=yes;

DATAROW=2;

RUN;

data Sales\_import;

set Sales\_import;

dummyType=(Type="R");\*Dummy variable -row homes home;

dummyRegion=(Region="M");\*Dummy variable -midwest;

RUN;

\*Scatterplot - Price \* Cost;

PROC GPLOT;

PLOT Price\*(Cost);

RUN;

\*Scatterplot - Price \* Home;

PROC GPLOT;

PLOT Price\*(dummyType);

RUN;

\*Scatterplot- Selling Price \* Region;

PROC GPLOT;

PLOT Price\*(dummyRegion);

RUN;

\*Correlation Values;

PROC CORR;

VAR dummyRegion dummyType Price Cost;

RUN;

\*Regression Analysis;

PROC REG;

MODEL Price=Cost dummyRegion dummyType;

RUN;

\*Regression Model w/o dummyRegion;

PROC REG;

MODEL Price=Cost dummyType;

RUN;

\*Regression model- Price;

PROC REG data=Price;

MODEL Price=Cost dummyType;

\*Residual Plot;

PLOT student.\*predicted.;

PLOT student.\*(Cost dummyType);

PLOT npp.\*student;

RUN;

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

The file banking.txt attached to this assignment contains the full dataset. 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 $

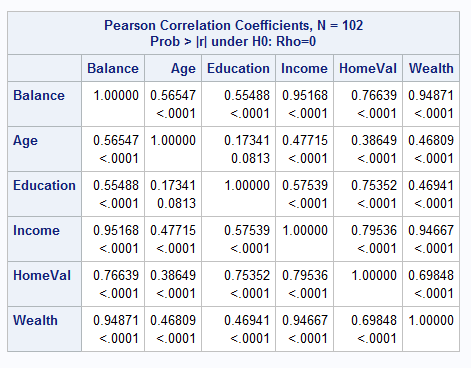
− median home value (HOMEVAL) in $

− median household wealth (WEALTH) in $

− average bank balance (BALANCE) in $

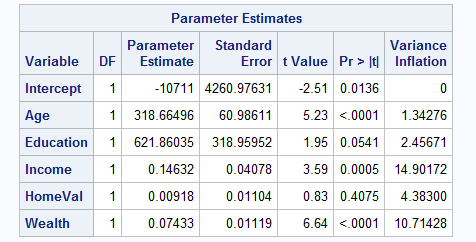
The goal of this exercise is to define a regression model to predict the average bank balance as a function of the other variables.

a) 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.

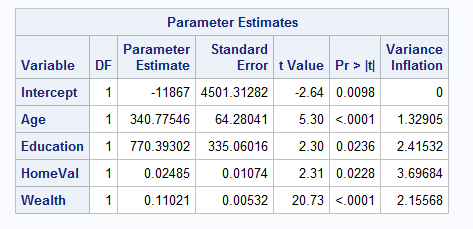


**The strongest correlation value for Balance is with Income. This is because the higher income you make the higher the balance you will have and vice versa. This same sentiment can be transcribed to the second strongest correlation value of Balance and Wealth. Home value has a slight correlation to balance. It does not have that strong of a correlation to balance because any income is able to save money to afford their dream home. Also loans exist for those who cannot afford a home at the moment but would like one. Age and Education do not have a correlation to balance. This is because anyone can get rich at random, for example, the lottery. The same logic can be correlated to education.**

b) 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. If you had taken an action for multicollinearity, fit the final model (Model M2). Include the outputs for both M1 and M2 models. Which model has the largest adj-R2 value?



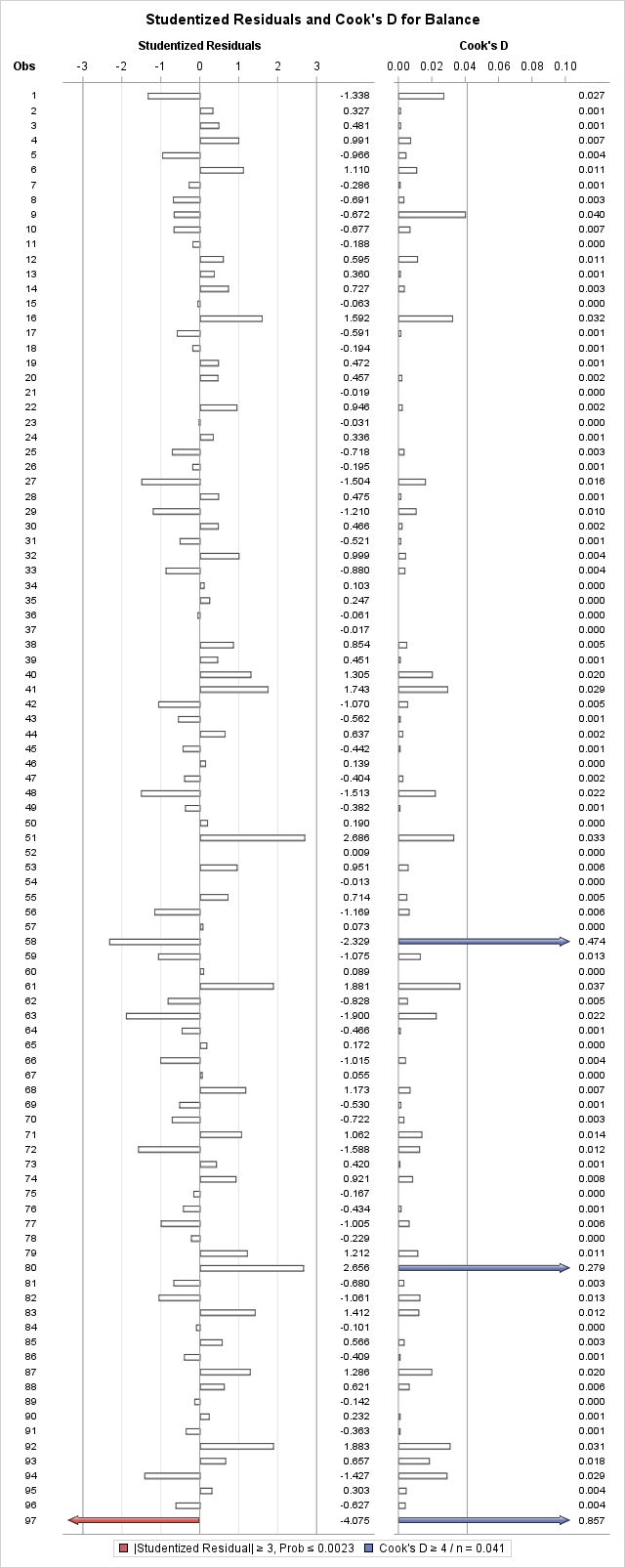
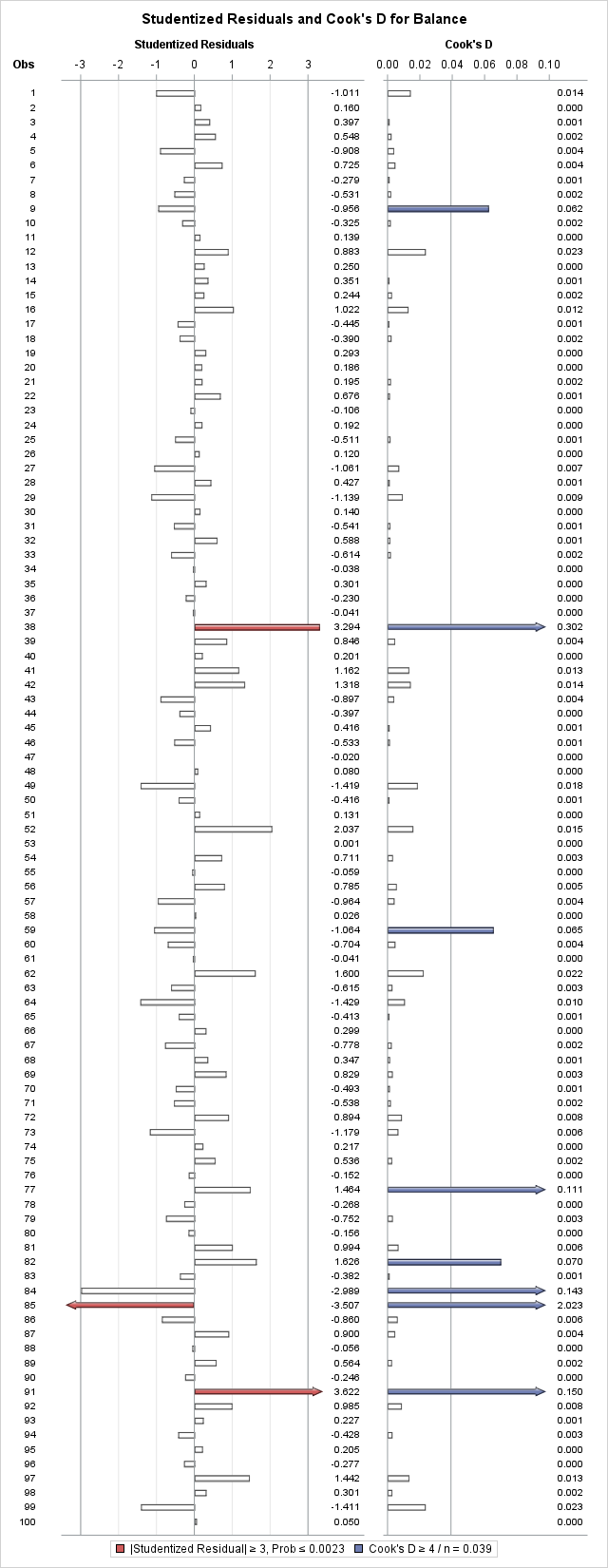
**Against the VIF of greater or equal to ten - Income and wealth come in at over ten. Income is the highest, therefore we will remove income.**

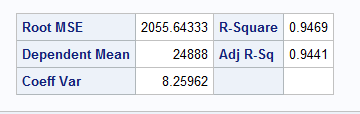
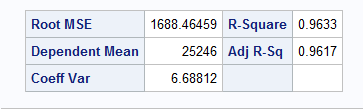
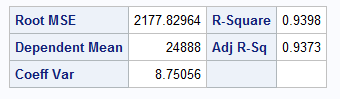


**Without income all VIF values are under ten. This means there is no collinearity.**

c) 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 outputs. Also rerun the final model, and report adj-R2 Value.

**There are five observations that are outliers: (1) #38 - positive influence, (2) #77 - positive influence, (3) #84 - positive influence, (4) #85 - negative influence, (5) #91 - Both influences. All five will be removed.**





**There are three more outlier points: (1) #58 - Positive influence, (2) #80 - Positive influence, and (3) #97 - negative influence. These will not be removed because of the Adjusted R-squared value raised by less than 3-5% and lowered. This means that the first outliers being removed creates the smallest spread.**

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

**PROC IMPORT DATAFILE="C:\Users\JDORETTI\Downloads\Banking.txt" out=banking replace;**

**DELIMITER='09'x;**

**GETNAMES=yes;**

**DATAROW=2;**

**RUN;**

**\*Correlation Values;**

**PROC CORR;**

**VAR Balance Age Education Income HomeVal Wealth;**

**RUN;**

**\*Regression - VIF;**

**PROC REG;**

**MODEL Balance=Age Education Income HomeVal Wealth/VIF;**

**RUN;**

**\*Regression- VIF w/o Income;**

**PROC REG;**

**MODEL Balance=Age Education HomeVal Wealth/VIF;**

**RUN;**

**\*Regression- Balance;**

**PROC REG;**

**MODEL Balance=Age Education HomeVal Wealth;**

**\*Residual Plot;**

**PLOT student.\*predicted.;**

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

**PLOT npp.\*student;**

**RUN;**

**\*Regression- no outlier;**

**PROC REG;**

**MODEL Balance=Age Education HomeVal Wealth/influence r;**

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

**PLOT npp.\*student.;**

**RUN;**

**\*no outliers;**

**data banking2;**

**set banking;**

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

**RUN;**

**\*Regression- no outliers;**

**PROC REG;**

**MODEL Balance=Age Education HomeVal Wealth;**

**RUN;**

**\*No outliers;**

**PROC reg data=banking2;**

**MODEL Balance=Age Education HomeVal Wealth/influence r;**

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

**PLOT npp.\*student.;**

**RUN;**

**\*no outliers;**

**data banking3;**

**set banking2;**

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

**RUN;**

**\*no outliers;**

**PROC REG;**

**MODEL Balance=Age Education HomeVal Wealth;**

**Problem 3 [10 pts]– to be answered by everyone**

Analytics is used in many different sports and has become popular with the Money Ball movie. The golf.csv dataset contains data about 196 tour players. 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="golf.csv" out=golf replace;

delimiter=',';

getnames=yes;

**run**;

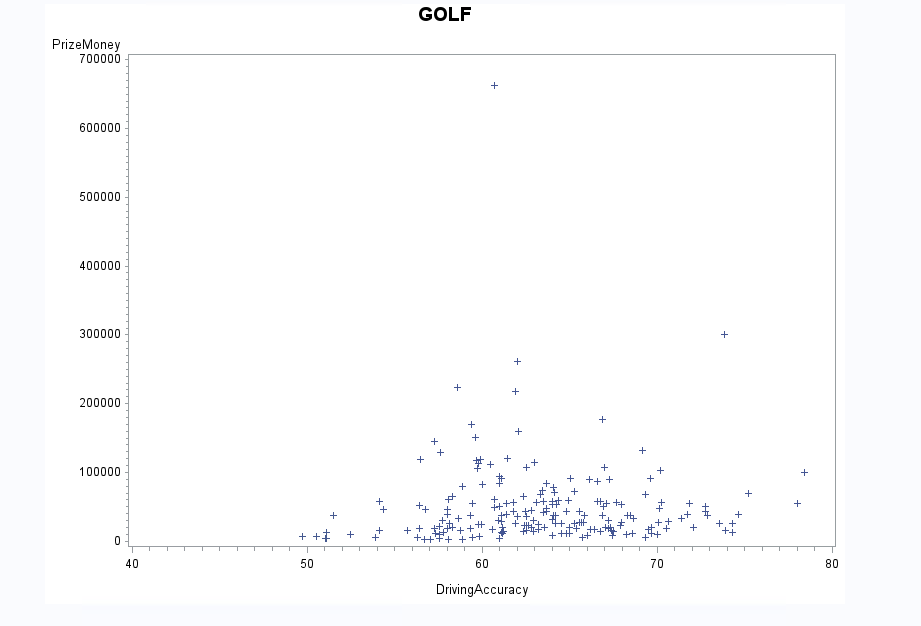
Note:

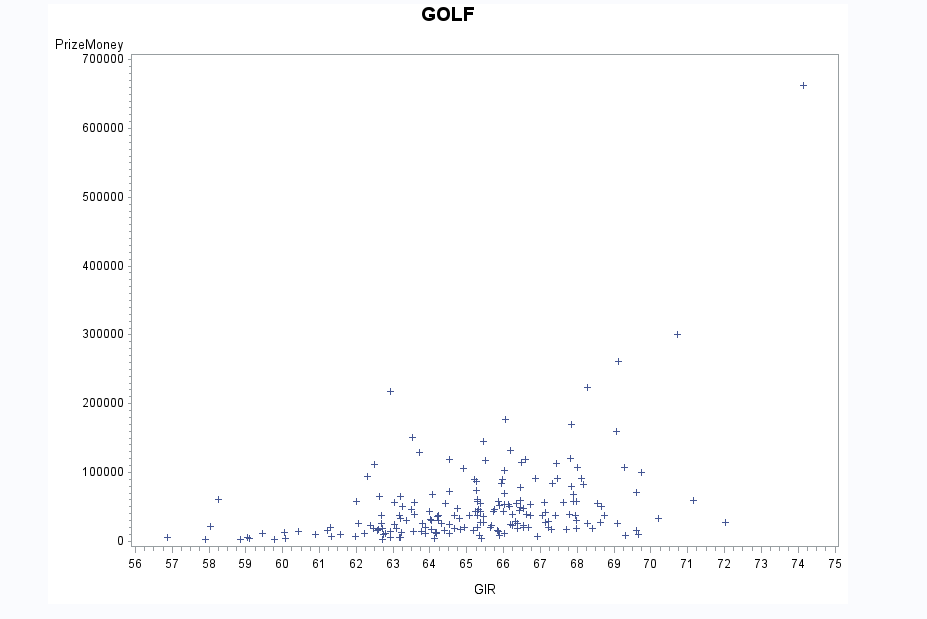
• The data file is in CSV format

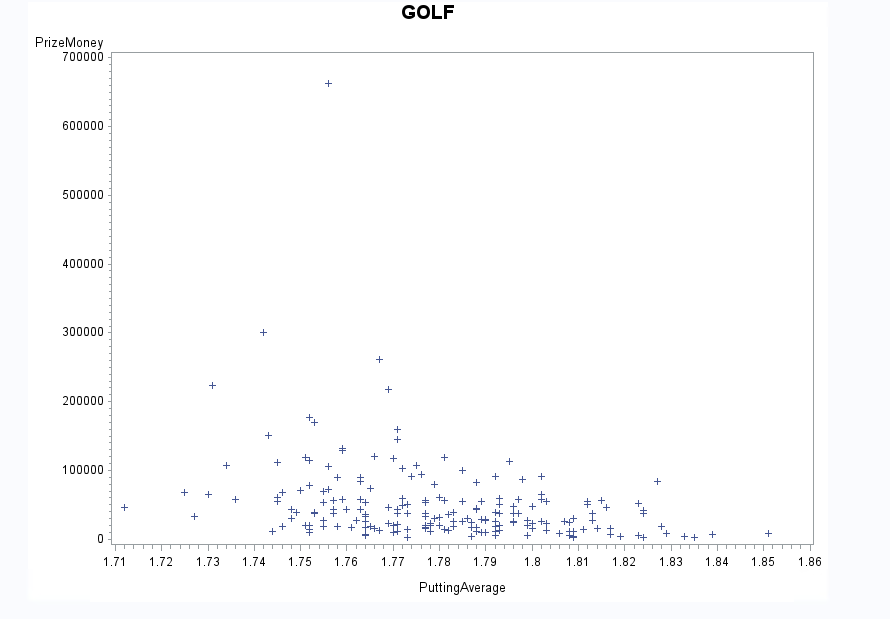
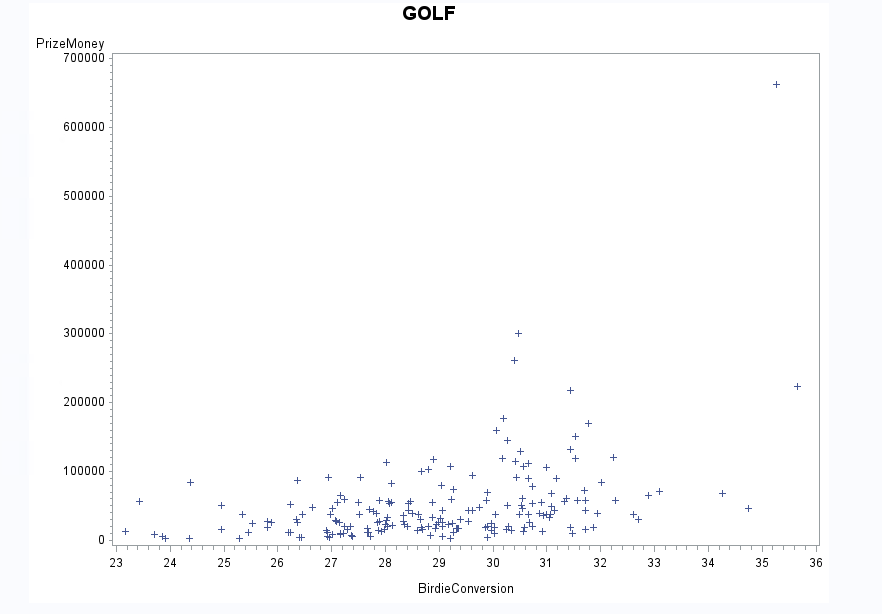
• It is delimitered with a comma

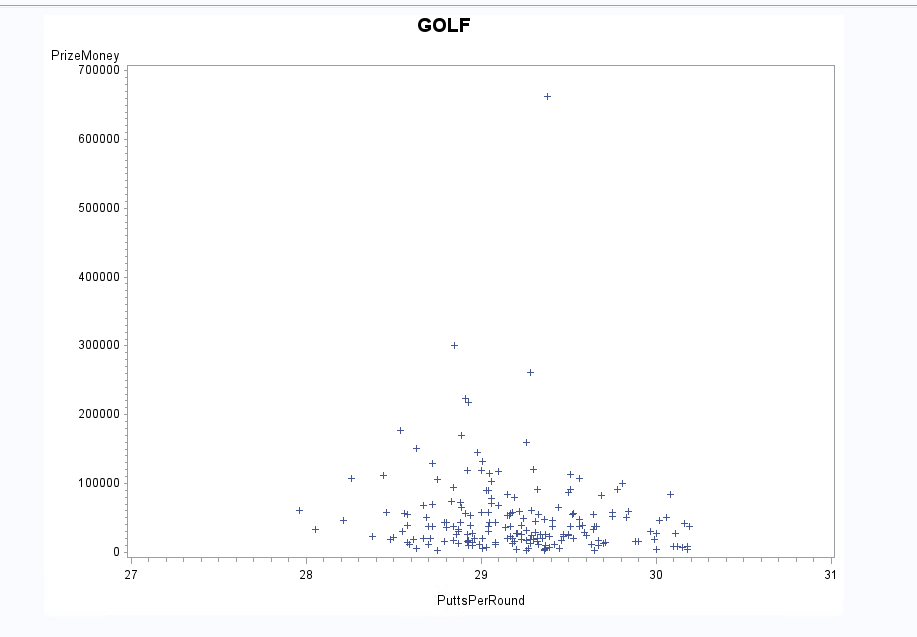
• The SAS dataset it is writing into is golf. You can change the name if you like.

a) 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.









**Price money to driving accuracy have no correlation. This is because prize money will always attract good talent, as top talent is likely to play in some leagues below them to have higher chances. Along with talent below and at the levels of the league. Resulting in a driving accuracy that is scattered and does not represent prize money.**

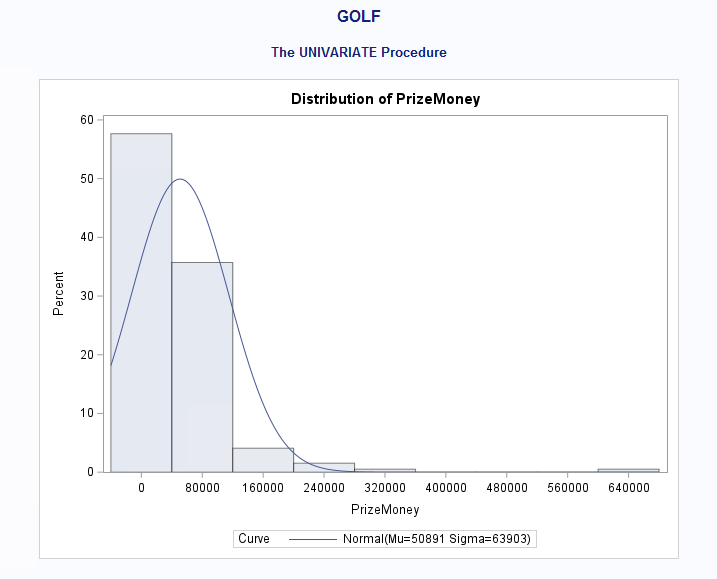
**Prize money to GIR has a very low positive correlation because similar to prize money to driving accuracy there are different levels of talent. However, adding up to two shots gives thoses with lower sill levels the opportunity to correct their swing and trajectory of the ball.**

**Prize money to Birdie Conversion also has no correlation. The spread is concentrated near the $100,000 level. Another issue with the data is that golf tournaments that are not professional may not have that high of prize pools. Skewing the data with outliers.**

**Prize money to putting average has a slight negative correlation. This is because with golf you want to hit the ball less times - therefore a lower putting score is better and that is why we see a slight correlation between higher prize money games and lower putting average.**

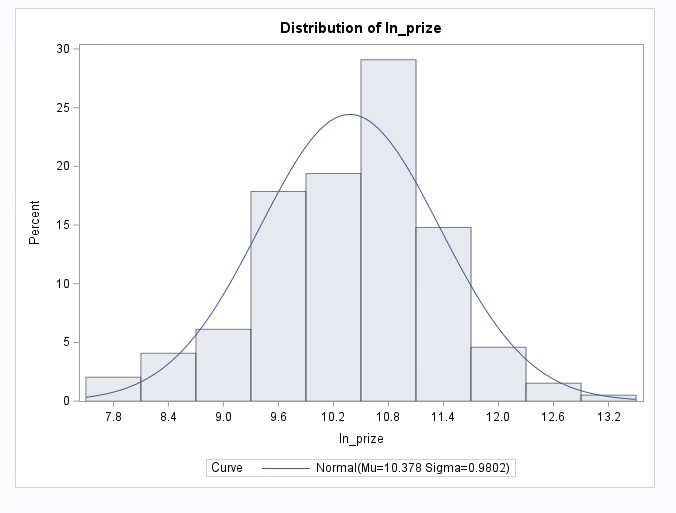
**Prize money and putts per round have no correlation once again because of different talent levels per event. Some players are better than the others.**

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



**Prize money’s histogram is positively skewed with a peak at $0. The outliers are $640,000 causing a right tail.**

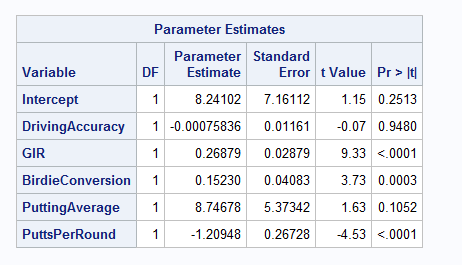
c) Apply a log transformation to PrizeMoney and compute the new variable ln\_Prize=log(PrizeMoney). Analyze distribution of ln\_Prize, and discuss if the distribution is symmetric or skewed. Include the relevant output.



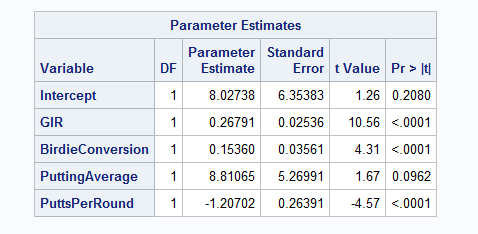
**This is a normal distribution with a single peak at 10.8. The only possible outlier could be 13.2 because there is a slight right tail.**

d) 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.

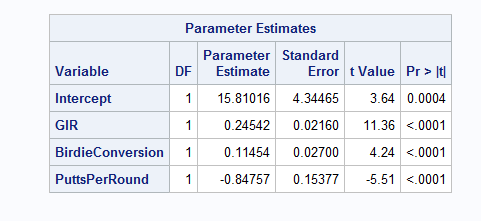
1) If necessary remove the non-significant variables. Remember to remove one variable at a time (variable with largest p-value is removed first) and refit the model, until all variables are significant.



**Here, driving accuracy has a p-value of .095 causing it to not be significant because the alpha test is less than .05.**

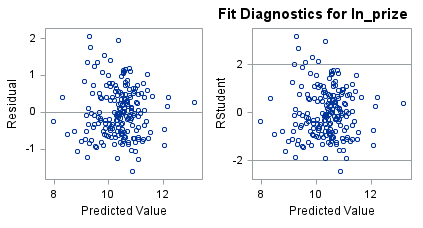


**Here, putting average has a p-value of .096 causing it to not be significant because the alpha test is less than .05.**

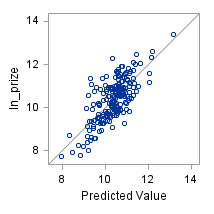


**Here, all p-values are less than .0001 meaning they are significant according to the alpha test value of less than .05.**

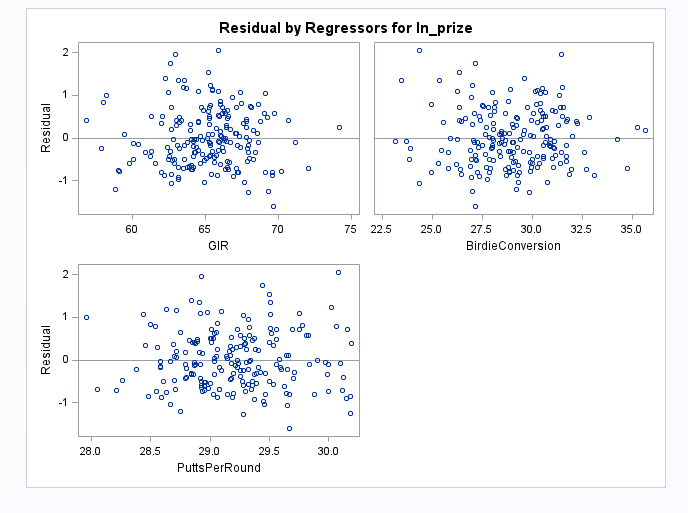
2) Analyze residual plots to check if the regression model is valid for your data. Discuss your analysis.



There are no patterns to be found in the above two plots. No pattern fits respectively into the plots.

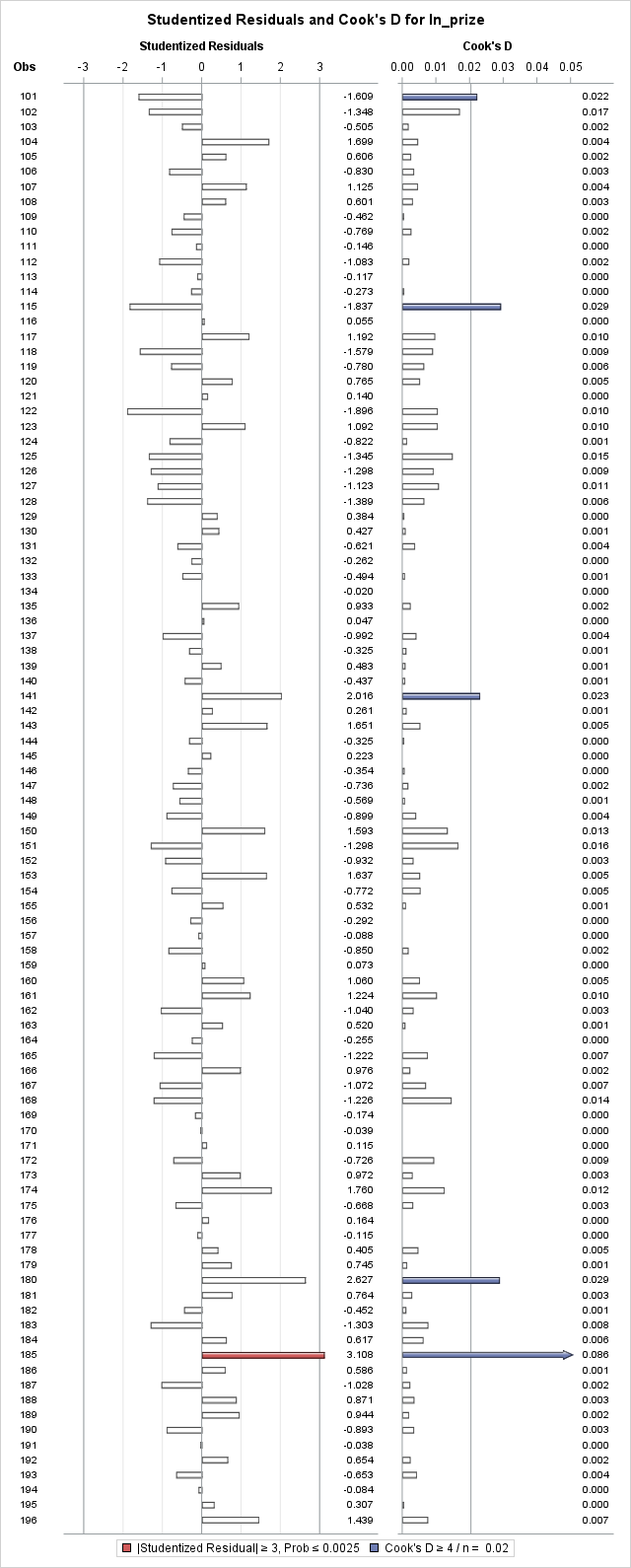
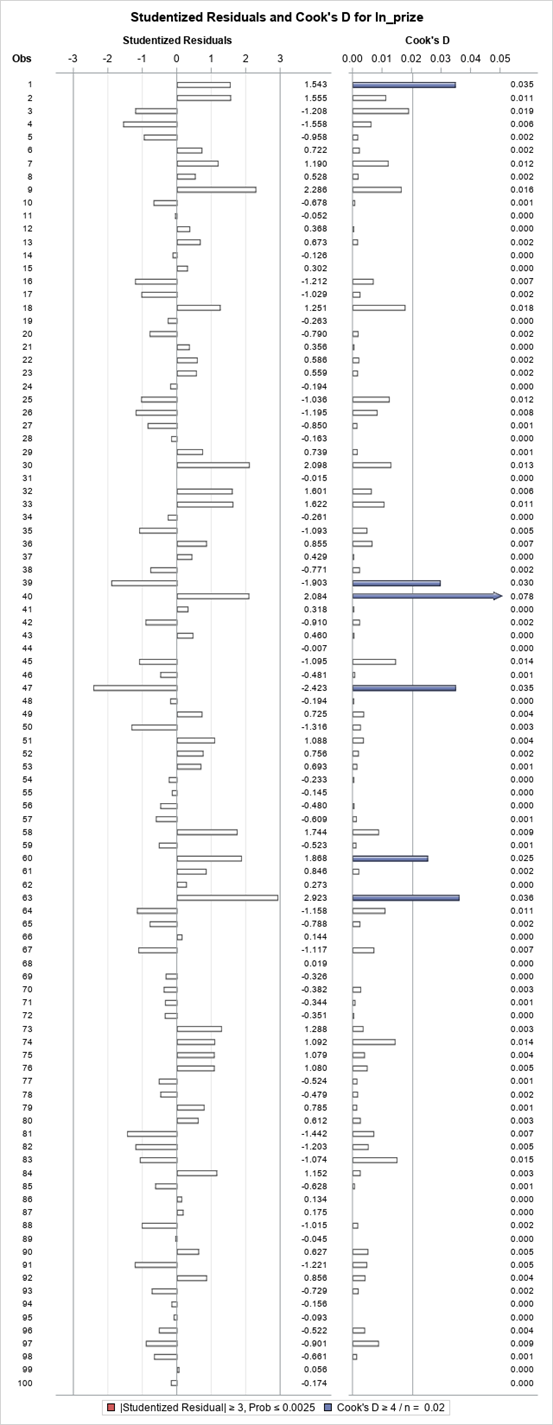


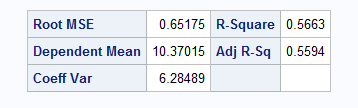
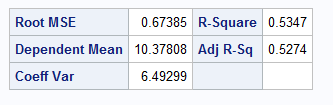
**This plots as a close spread shown by the dark blue dots on the plot. This happens because of the concentration of the data points. The graph is also positively linier.**



**All three plots also have no patterns. No patterns fit on these plots repecitlby, therefore causing these plots to have no pattern.**

3) 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.

**There are two outliers that need to be removed: #40 - positive influence and #185 - positive influence. These will be removed.******



**Based on the Adjusted R-Squared and the insignificant change along with no more arrows in the second running of the code - all outliers were removed.**

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

**Model equation is the best model - rounded by .5. In\_prize = 15.81 + .25(GIR) + .12 (Birdie conversion) - .85 (puttsperround)**

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

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

delimiter=',';

getnames=yes;

run;

TITLE"GOLF";

PROC PRINT;

RUN;

\*ScatterPlot - PrizeMoney \* DrivingAccuracy;

PROC GPLOT;

PLOT PrizeMoney\*(DrivingAccuracy);

RUN;

\*ScatterPlot- PrizeMoney \* GIR;

PROC GPLOT;

PLOT PrizeMoney\*(GIR);

RUN;

\*ScatterPlot- PrizeMoney \* BirdieConversion;

PROC GPLOT;

PLOT PrizeMoney\*(BirdieConversion);

RUN;

\*ScatterPlot- PrizeMoney \* PuttingAverage;

PROC GPLOT;

PLOT PrizeMoney\*(PuttingAverage);

RUN;

\*ScatterPlot- PrizeMoney \_ PuttsPerRound;

PROC GPLOT;

PLOT PrizeMoney\*(PuttsPerRound);

RUN;

\*Histogram;

PROC UNIVARIATE normal;

VAR PrizeMoney;

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

RUN;

\*log prize;

data golfGame;

set golf;

ln\_prize=log(PrizeMoney);

PROC PRINT;

RUN;

\*Histogram;

PROC UNIVARIATE normal;

VAR ln\_prize;

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

RUN;

\*Regression- ln\_prize;

PROC REG;

MODEL ln\_prize=DrivingAccuracy GIR BirdieConversion PuttingAverage PuttsPerRound;

RUN;

\*Regression- ln\_prize w/o DrivingAccuracy;

PROC REG;

MODEL ln\_prize=GIR BirdieConversion PuttingAverage PuttsPerRound;

RUN;

\*Regression- ln\_prize w/o PuttingAverage;

PROC REG;

MODEL ln\_prize=GIR BirdieConversion PuttsPerRound;

RUN;

\*Residual- ln\_prize;

PROC REG;

MODEL ln\_prize=GIR BirdieConversion PuttsPerRound;

PLOT student.\*predicted.;

PLOT student.\*(GIR BirdieConversion PuttsPerRound);

PLOT npp.\*student;

RUN;

PROC REG;

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

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

plot npp.\*student.;

RUN;

\*No outliers;

data golfGame2;

set golfGame;

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

RUN;

\*Regression- ln\_prize w/o PuttingAverage;

PROC REG;

MODEL ln\_prize=GIR BirdieConversion PuttsPerRound;

RUN;

\*No outliers;

PROC REG;

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

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

plot npp.\*student.;

RUN;