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

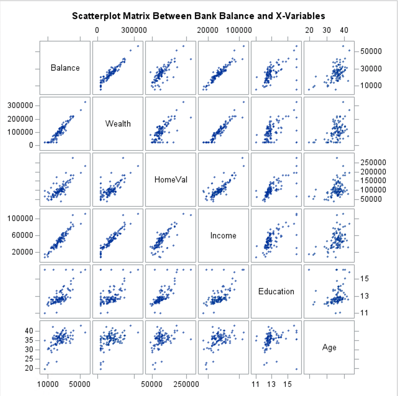
The file banking.txt attached to this assignment contains the full dataset. It provides data acquired from banking and census records for different zip codes in the bank’s current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices.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

* + - 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)

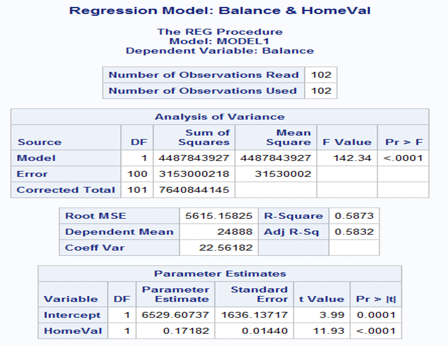
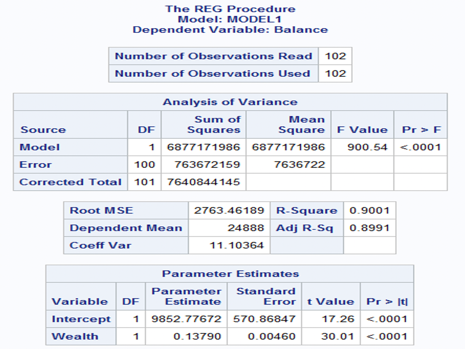
Matrix Plot

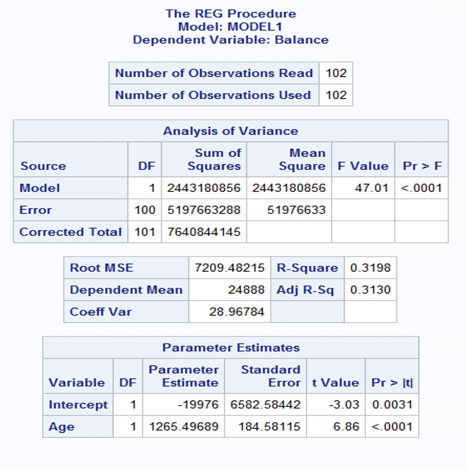
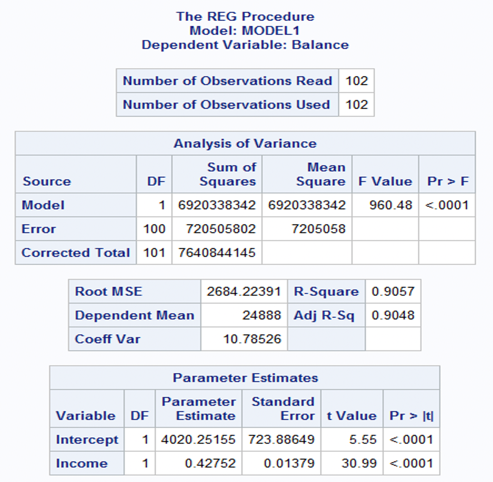


**Discussion:** Matrix Plots that are outlines in green (Balance & Wealth, Balance & HomeVal, Balance & Income, HomeVal & Income) appear to have an overall linear appearance. All plots that are not outlined are suspected to be non-linear due to what appears to be potential outliers and pattern-like appearance. If you observe the suspected non-linear scatter plots (non-outlined plots) you will notice plot points deviating upwards throughout the length of dependent variables (for example, Education). Otherwise, if you consider other dependent variables such as Age (also seen in Education, Income, HomeVal, and Wealth) you will notice a clustering resembling a logarithmic growth or S-like appearance.

Regression Analysis:

Table

Description automatically generated



b) 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:**

Balance & Wealth: r = ~0.9487 **Serious Correlation Problem**

Balance & HomeVal: r = ~0.7663 **Moderate-High Correlation**

Balance & Income: r = ~0.9516  **Serious Correlation Problem**

Balance & Age: r = ~0.5655 **Moderate Correlation**

Balance & Education: r = ~0.5548 **Moderate Correlation**

**Discussion:** The most correlated x-variables from greatest to least are Income, Wealth, HomeVal, Age then Education. All variables are positively correlated ranging from moderate to (very) high correlation.

* + - 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.

**Regression M1 with VIF Statistics & Tolerance Value**

**Table

Description automatically generatedTable

Description automatically generated**

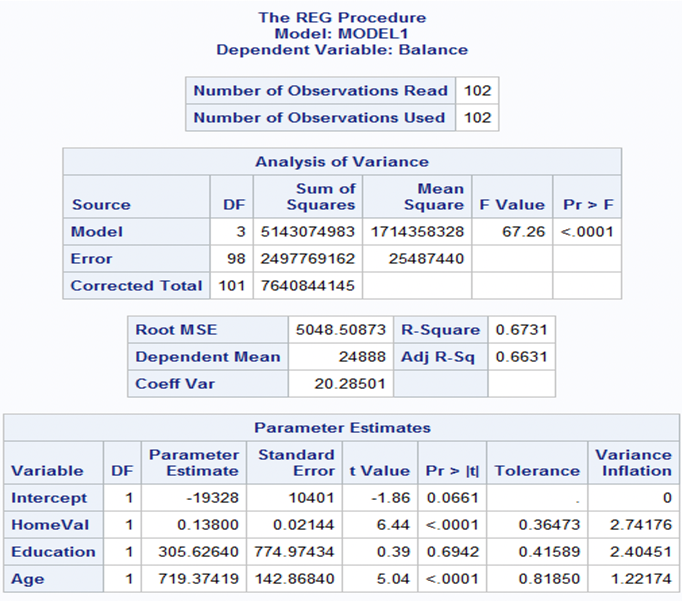
**Note: Tol <= 0.10 suggests collinearity. VIF >= 10.0 suggests collinearity**

**Discussion:** When we consider the tolerance values (tol) and variance inflation factor (vif) statistics of x-variables within the banking data set we will notice that variables Wealth and Income for each multicollinearity diagnostic technique suggest collinearity such that Wealth and Income have a tolerance value less than 0.10 and a vif statistic greater than 10.0. Overall, both the tolerance values and vif statistics suggest collinearity. Since multicollinearity can reduce precision, it is advised to change the model by dropping these variables of obtain more data.

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

**Regression M2 with VIF Statistics & Tolerance Value Compared to the above M1**

(After removing Wealth & Income) (After removing Wealth & Income & Education)

Table

Description automatically generated

* 1. Analyze the adj-R2 values for both models M1 and M2. Which model has the largest adj-R2 value?
     1. M1: : 0.9441
     2. M2: : 0.6660
     3. Discussion: M1 has a greater at 0.9441.
  2. 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.

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Discussion: Considering the above graphs, we can recognize that the predicted and studentized residual plots maintain an equal independence, variance, or random scatter around the zero-line. However, the normal plot does not appear linear but S-shaped, a potential solution would be to transform the data.**

3)  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.

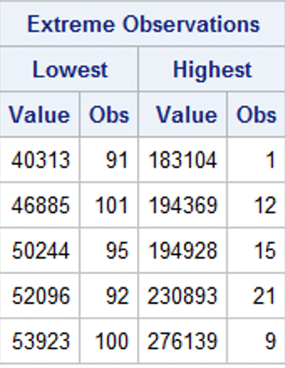
**Histogram:**

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Table

Description automatically generated with medium confidence

**Box Plot:**

**Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated**

**Percentile:**

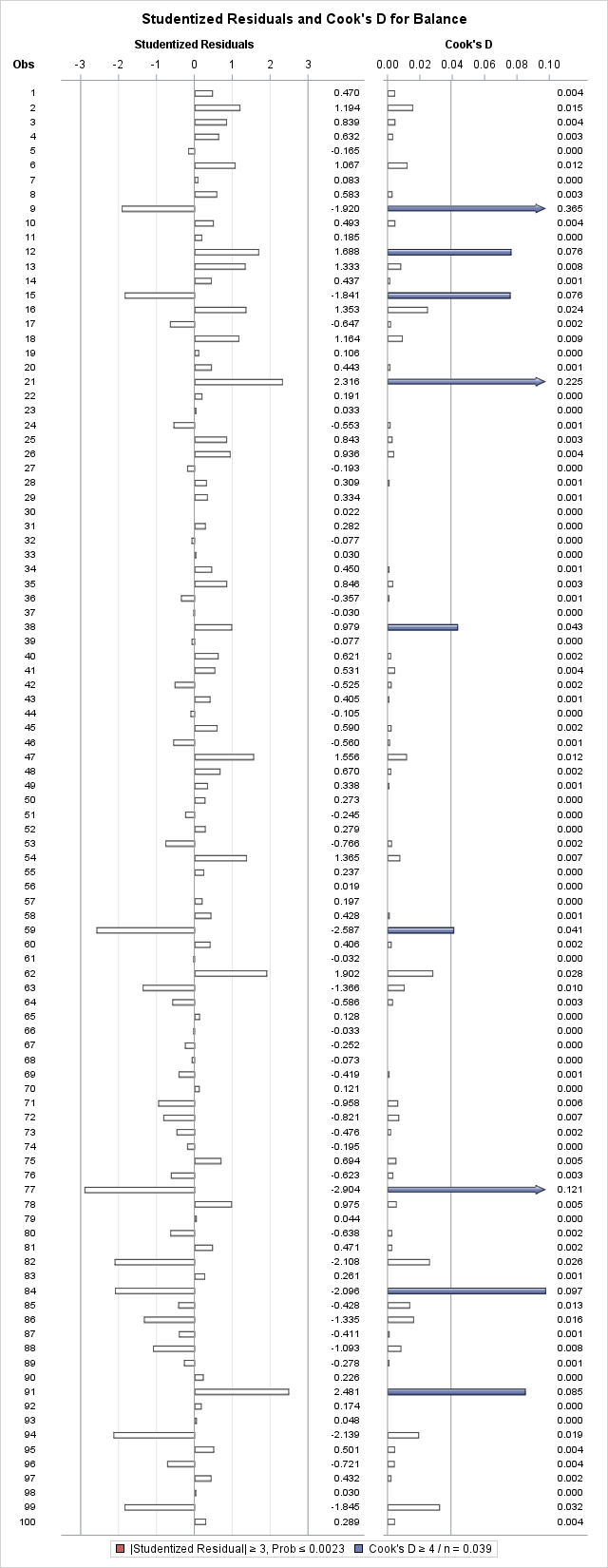
**Table

Description automatically generatedTable

Description automatically generated**

**Discussion:** Considering the graphs and tables above, we notice that there does exist outliers and influential points. Starting with the histogram data, we notice there are no observations concerning 27-year-old individuals and no observations for home values at 255,000$ which may suggest that observations of individuals who are 21- and 24-year-old and home values beyond 255,000$ are outliers. In regard to the extreme observations, we become aware of outliers, fringeliers, and influential points. Similarly, when considering the boxplot, we notice outliers within similar value ranges as seen in our histograms. Overall, now that we aware of our outliers we can choose to either note that they exist or consider removing them from our observations. We can determine observation removal via influential point detection statistics (leverage statistics hii, cov ratio, dffits, dfbetas, cook’s distance).

**Computing Studentized Residuals and Cook D**



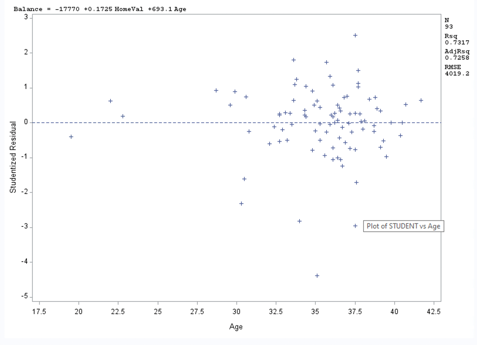
Considering the studentized residuals and Cook D graph, we notice that there are influential points at observation 9, 12, 15, 21, 38, 59, 77, 84 and 91 so we have removed them to improve our data set. However in our new data set banking\_new we discover more influential points and outliers. I removed those observations (93, 90, 85, 75, 56) to improve the data set a little more.

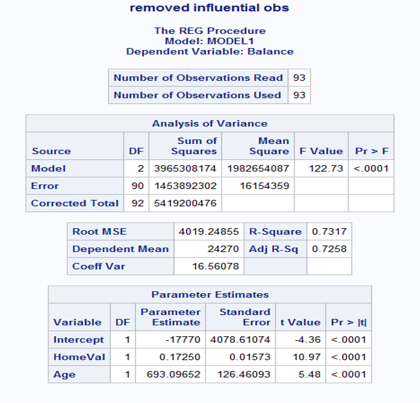
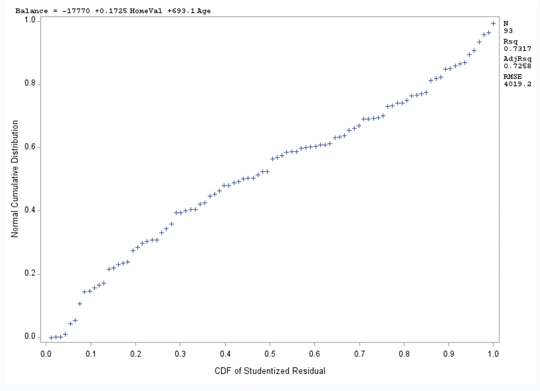
**Removing influential observations**

**Banking\_new**

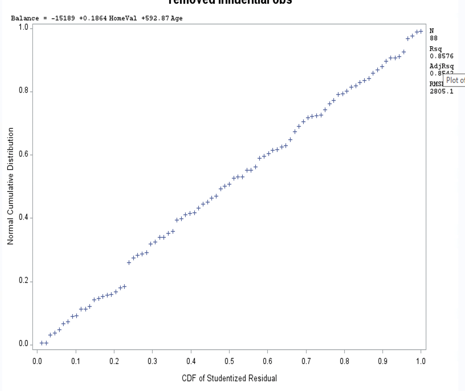
**Chart, scatter chart

Description automatically generatedChart, scatter chart

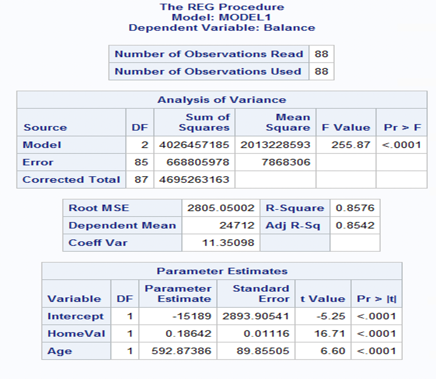
Description automatically generated**

****

**Banking\_new\_2**

**Chart, scatter chart

Description automatically generated**

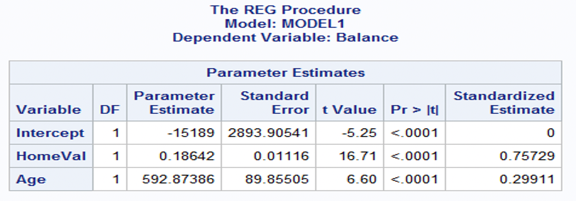
**Chart, scatter chart

Description automatically generated**

**Discussion:** After removing the influential points and outliers a second time we get a greater F-Value (Banking\_new: 122.73 vs Banking\_new\_2: **255.87**) and a greater Adj-R2 (Banking new: 0.7258 vs. Banking\_new\_2: **0.8542**) which further proves a better to fit to the model.

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

**M2:Banking\_New\_2 Standardized Coefficients**

****

**Discussion:** HomeVal has the strongest influence on balance.

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

**proc** **import** file = "C:\Users\DVALDE12\Desktop\A4\Banking.txt"

out = banking replace;

**run**;

\*matrix plot;

**proc** **sgscatter**;

title "Scatterplot Matrix for all X-var";

matrix Balance Wealth HomeVal Income Education Age;

**run**;

\*All X Var;

**proc** **reg** data = banking;

title "Correlation Between Balance and All X-Var";

model Balance= Wealth HomeVal Income Education Age;

**run**;

\*Wealth;

**proc** **reg** data = banking;

title "Correlation Between Balance and Wealth";

model Balance= Wealth;

**run**;

\*HomeVal;

**proc** **reg** data = banking;

title "Correlation Between Balance and HomeVal";

model Balance=HomeVal;

**run**;

\*Income;

**proc** **reg** data = banking;

title "Correlation Between Balance and Income";

model Balance=Income;

**run**;

\*Education;

**proc** **reg** data = banking;

title "Correlation Between Balance and Edu";

model Balance=Education;

**run**;

\*Age;

**proc** **reg** data = banking;

title "Correlation Between Balance and Age";

model Balance=Age;

**run**;

\*vif and multicollinearity;

**proc** **reg** data = banking;

title "Correlation Between Balance and All X-Var";

model Balance= Wealth HomeVal Income Education Age / vif;

**run**;

\*tolerance and multicollinearity;

**proc** **reg** data = banking;

title "Tolerance Between Balance and All X-Var";

model Balance= Wealth HomeVal Income Education Age / tol;

**run**;

\*Histogram;

**proc** **univariate** data = banking;

var HomeVal Age;

histogram;

**run**;

\*Boxplot;

**proc** **sgplot** data = banking;

vbox Balance;

**run**;

**proc** **sgplot** data = banking;

vbox HomeVal;

**run**;

**proc** **sgplot** data = banking;

vbox Age;

**run**;

\*percentile;

**proc** **univariate** data = banking;

var HomeVal;

output out = percentile\_data

pctlpts = **0** **1** **5** **10** **25** **50** **75** **90** **95** **99** **100**

pctlpre = P\_;

**run**;

**proc** **univariate** data = banking;

var Age;

output out = percentile\_data

pctlpts = **0** **1** **5** **10** **25** **50** **75** **90** **95** **99** **100**

pctlpre = P\_;

**run**;

\*Cook D & Studentized residuals;

**proc** **reg** data = banking;

model Balance = HomeVal Age / influence r;

plot student.\*(HomeVal Age predicted.);

plot npp.\*student.;

**run**;

\*Removing influential points;

**data** banking\_new;

set banking;

if \_n\_ in (**9**, **12**, **15**, **21**, **38**, **59**, **77**, **84**, **91**) then delete;

**run**;

\*new model;

**proc** **reg** data = banking\_new;

title "removed influential obs";

model Balance = HomeVal Age / influence r;

plot student.\*(HomeVal Age predicted.);

plot npp.\*student.;

**run**;

\*removing some more points;

**data** banking\_new\_2;

set banking\_new;

if \_n\_ in (**93**, **90**, **85**, **75**, **56**) then delete;

**run**;

\*another new model;

**proc** **reg** data = banking\_new\_2;

model Balance = HomeVal Age / influence r;

plot student.\*(HomeVal Age predicted.);

plot npp.\*student.;

**run**;

\*standardized coefficients;

**proc** **reg** data = banking\_new\_2;

model Balance = HomeVal Age / stb;

ods select ParameterEstimates;

**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 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.

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 Matrix**

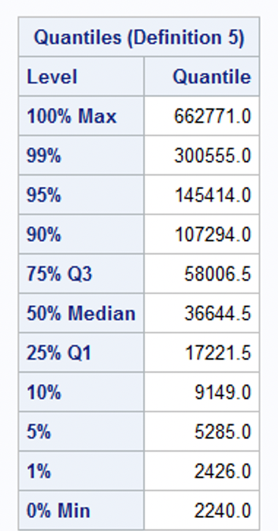
Diagram

Description automatically generated with medium confidence

**Discussion:** Considering the x-variables in relation to the y-variable (PrizeMoney) in the matrix, there appears to be non-linear.

1. Analyze distribution of PrizeMoney and discuss if the distribution is symmetric or skewed. Include the relevant output.
   1. **Distribution Analysis:**

**Table

Description automatically generated**

**Discussion:** Considering the distribution analysis, we have a skewness of approximately 5.4 which signifies significant skewing to the right.

* 1. **Histogram:**

Chart, histogram

Description automatically generated

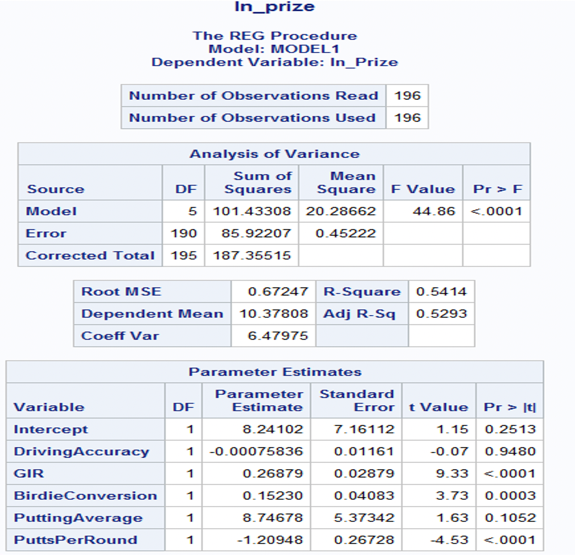
**Discussion:** Considering the Histogram we know that PrizeMoney is right skewed.

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.

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.

**First regression Model DrivingAccuracy Removed DrivingAccuracy & PuttingAvg. Removed**

Table

Description automatically generatedTable

Description automatically generated

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

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**Discussion:** Considering the above graphs, there seems to be consistency in independence, variance, and linearity. Therefore, the residual plots suggests that the regression model is suitable for the data

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.

A picture containing text, device, thermometer, measuring stick

Description automatically generatedA picture containing text, device, thermometer

Description automatically generated

**Discussion:** Considering our studentized residual and Cook D graph, we know that outliers and influential points exist in our dataset. It is important to remove particularly extreme or influential observations since these data points can powerfully mislead the true nature of our data and misrepresent our analysis.

**We have an outlier at observation 185 and influential points at 1, 38, 39, 40, 47, 60, 63, 101, 114, 141, 180, 185**

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Discussion: By removing outliers and influential points, we can ensure model accuracy.**

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

Table

Description automatically generated

**Fitted Model: Ln(PrizeMoney)** = $14.60003 + $0.26731**GIR** + $0.12523**BirdieCoversion** - $0.86681**PuttsPerRound**

**Discussion:** The fitted model is the best model since we have removed x-variables with p-values greater than 0.05 and removed outliers/influential observations that manipulate the overall data. Overall, creating a true linear, non-skewed transformation of our original dataset with a supportive goodness-of-fit (f-value: 102.22, Adj-R2: 0.6240, R2: 0.6301, R: ~.0.79).

e)  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?

**Discussion:** Since GIR is a significant value in the fitted model, increasing GIR by 1% will increase the amount of PrizeMoney.

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

**proc** **import** file = "C:\Users\DVALDE12\Desktop\A4\golf.csv"

out = Golf

dbms = csv

replace;

**run**;

**proc** **print** data = Golf;

**Run**;

**proc** **sgscatter**;

title "Scatterplot Matrix for PrizeMoney & all x-var";

matrix PrizeMoney DrivingAccuracy GIR BirdieConversion PuttingAverage PuttsPerRound;

**run**;

\*distributive analysis;

**proc** **univariate** data = Golf;

var PrizeMoney;

**run**;

\*histogram;

**proc** **univariate** data = Golf;

var PrizeMoney;

Histogram;

**run**;

\*log PrizeMoney;

**data** Golf\_ln;

set Golf;

ln\_Prize = log(PrizeMoney);

**proc** **print** data = Golf\_ln;

var PrizeMoney DrivingAccuracy GIR BirdieConversion PuttingAverage PuttsPerRound ln\_Prize;

title "ln\_prize";

**run**;

\*removing largest p;

**proc** **reg** data = Golf\_ln;

model ln\_prize = DrivingAccuracy GIR BirdieConversion PuttingAverage PuttsPerRound;

**run**;

\*removed Driving Accuracy;

**proc** **reg** data = Golf\_ln;

model ln\_prize = GIR BirdieConversion PuttingAverage PuttsPerRound;

**run**;

\*removed PuttingAverage;

**proc** **reg** data = Golf\_ln;

model ln\_prize = GIR BirdieConversion PuttsPerRound;

**run**;

\*finding outliers and influential pts;

**proc** **reg** data =Golf\_ln;

title "outliers & influential obs ln\_Prize";

model ln\_Prize = GIR BirdieConversion PuttsPerRound / influence r;

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

plot npp.\*student.;

**run**;

\*removing outliers;

**data** Golf\_ln\_2;

set Golf\_ln;

if \_n\_ in (**1**, **38**, **39**, **40**, **47**, **60**, **63**, **101**, **114**, **141**, **180**, **185**) then delete;

**run**;

\*new model;

**proc** **reg** data = Golf\_ln\_2;

title "removed influential obs";

model ln\_Prize = GIR BirdieConversion PuttsPerRound / influence r;

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

plot npp.\*student.;

**run**;

\*fitted model;

**proc** **reg** data = Golf\_ln\_2;

title "Fitted Model";

model ln\_prize = GIR BirdieConversion PuttsPerRound / stb;

**run**;