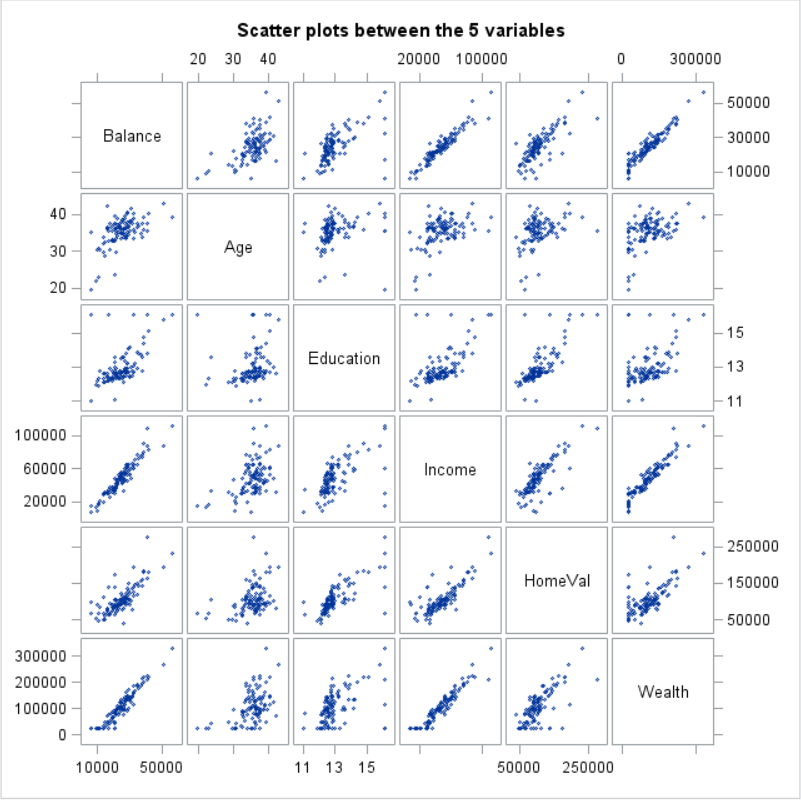
CSC 423 – ASSIGNMENT 4

BY

Sasidhar Mukthinuthalapati

1. Problem -1
   1. Scatterplots to visualize the associations between bank balance and the other five variables.



From the scatterplot matrix, we can make the following observations:

* Balance and Age: The correlation between the balance and age doesn’t seem to be linear but rather concentrated at a certain age group mainly between Age 30 – 45. For more information, we would need to refer to the correlation values.
* Balance and Education: The **correlation** between balance and education seems to be **somewhat linear** but **positive**. The strength of the correlation cannot be determined using these plots and hence we would need to study the correlation values.
* Balance and Income: Balance and Income are **strongly positively correlated** to each other. This means that with increase in Income there is a positive increase in balance.
* Balance and HomeVal: These two variables seem to be positively correlated but to arrive at a proper conclusion we need to refer to the correlation values. But the two variables are positively correlated.
* Balance and Wealth: These two variables are also strongly and positively correlated to each other. The strength of correlation cannot be determined using this plot though. Wealth and Income seem to be positively and strongly correlated to the Balance but the strongest contributor cannot be determined using these plots.

The Code to generate the scatter plot matrix is as follows:

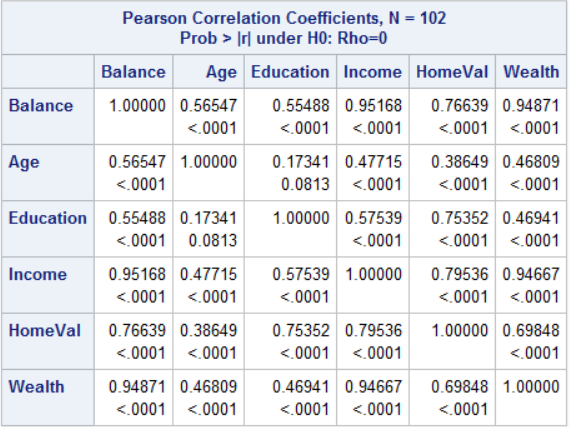
title "Scatter plots between the 5 variables";

**proc** **sgscatter**;

MATRIX balance age education income homeval wealth;

**run**;

* 1. The Pearson’s Correlation values between the variables are as follows:



The following observations can be made using the above values for the variables involved:

* Balance and Age: The correlation between balance and age is **positive and strong and linear** but not the strongest. With an increase in age there is a positive increase in balance. But with age we can only explain **56.54% of the variance in Balance**.
* Balance and Education: Education is the **weakest contributor among all the other variables** which are correlated to Balance. The **correlation though is positive and somewhat linear** but it is not as strong as other variables. Using education, we can only be able to explain **55.48% of the variance** in balance.
* Balance and Income: These are **strongly, positively and linearly correlated** to each other. Using Income, we can explain **95.16% of the variance in Balance**. Based on the correlation values we can say that Income is the strongest contributor of all.
* Balance and HomeVal: These two variables are **positively and linearly** correlated to each other and the **correlation is also strong but not as strong as Income and Wealth**. Using HomeVal, we can explain **76.63% of the variance** in balance.
* Balance and Wealth: These two variables are **strongly positively and linearly correlated** to each other. Using Wealth, we can explain up to **94.87% of the variance** in Balance. Wealth is the second-best contributor to predicting balance after Income.

Based on the above observation we can say the Balance has the strongest association to the Income and then Wealth and HomeVal and so on.

The code to generate the above-mentioned Pearson Correlation Values is the following:

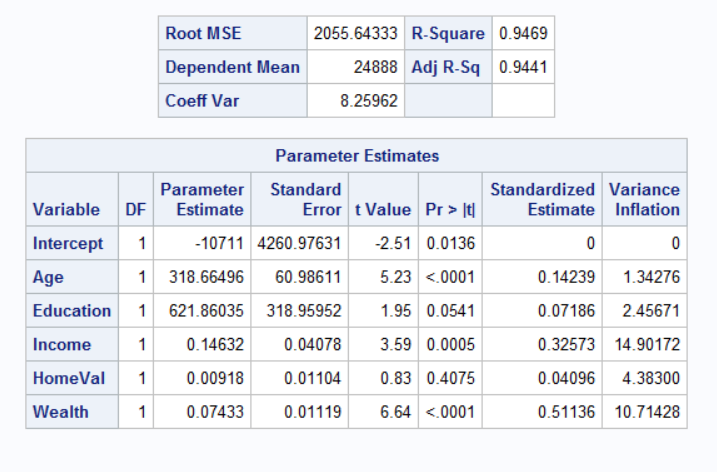
title "Pearson's Correlation Values";

**proc** **corr**;

var balance age education income homeval wealth;

**run**;

* 1. The Regression Model M1 which has all the five variables.



We can make the following observations from the above Output:

* In the above output, we have added the standard estimate and VIF (Variance Inflation) which is used to show if multicollinearity exists or not.
* In this is model we can see that multicollinearity exists because the VIF value is greater than 10 for 2 variables i.e. Income and Wealth.
* This claim can be further supported by referring to the Pearson’s Coefficient values table. In that table, we can see that Income and Wealth have a correlation of 94.66% which further explains the presence of near perfect collinearity in this model.
* Based on the above results we can either have Wealth or Income in our model.
* Also from the above output we can say that Education and HomeVal are not significant contributors because they have p-value which is greater than 0.05.

The code to generate the above output is as follows:

title "M1 model which has all the 5 variables and vif values";

**proc** **reg**;

model balance = age education income homeval wealth/ vif stb;

**run**;

1. The first model M1 suffers with multi collinearity because it has two Independent variables which are correlated to each other and in a strong manner. These variables are Wealth and Income and hence to come up with a new model I’ve decided to remove the multi collinearity, choose the model with highest adjusted R-Squared Value and remove the insignificant variables.

**Model 2:** In this model, I’ve removed the multicollinearity by removing the Income variable. In this model, education and HomeVal also seem to be significant contributors.

**Model 3:** In this model, I’ve removed the multicollinearity by removing the Wealth variable but have kept the insignificant variable HomeVal and Education.

**Model 4:** In this model, I’ve removed the insignificant variable as well as multicollinearity by removing wealth.

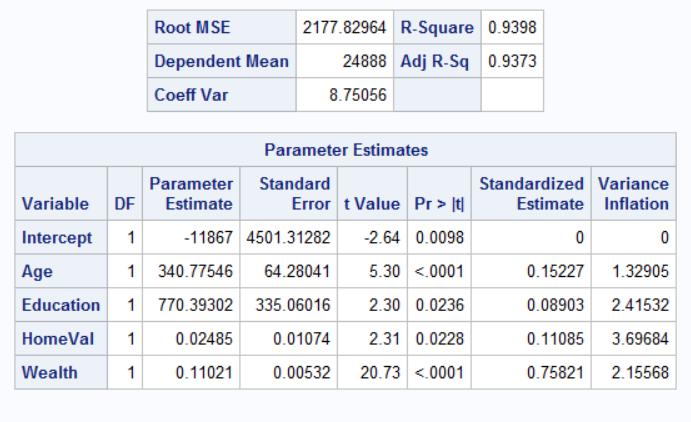
**Model 5:** In this model, I’ve replaced Wealth in model 4 with Income and removed the insignificant variables.

**Model 6:** In this model, I’ve removed the insignificant but have kept the income and wealth variable meaning multicollinearity still exists with respect to model 1. But when we compare the VIF values in this model it is still high and hence I didn’t consider this model to be my final model.

The above points can be summarized with the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R-Squared | Adjusted R-Squared | Non-Significant Variables | Multi Collinearity |
| M1 | .9469 | .9441 | 2 | YES |
| Model 2 | .9398 | .9373 | 0 | NO |
| Model 3 | .9225 | .9193 | 2 | NO |
| Model 4 | .9189 | .9173 | 0 | NO |
| Model 5 | .9218 | .9202 | 0 | NO |
| Model 6 | .9417 | .9399 | 0 | YES |

M2 would be Model 2 in my case. In this model we get the following parameter estimates values:



As we can see the p-value associated with education and HomeVal are less than 0.05 which means they can be considered as significant varaibles.

The code to generate the above model is as follows:

title "M2 model without income";

**proc** **reg**;

model balance = age education homeval wealth/ vif stb;

**run**;

We can see that Wealth is the strongest contributor because it has the highest standardized estimate value.

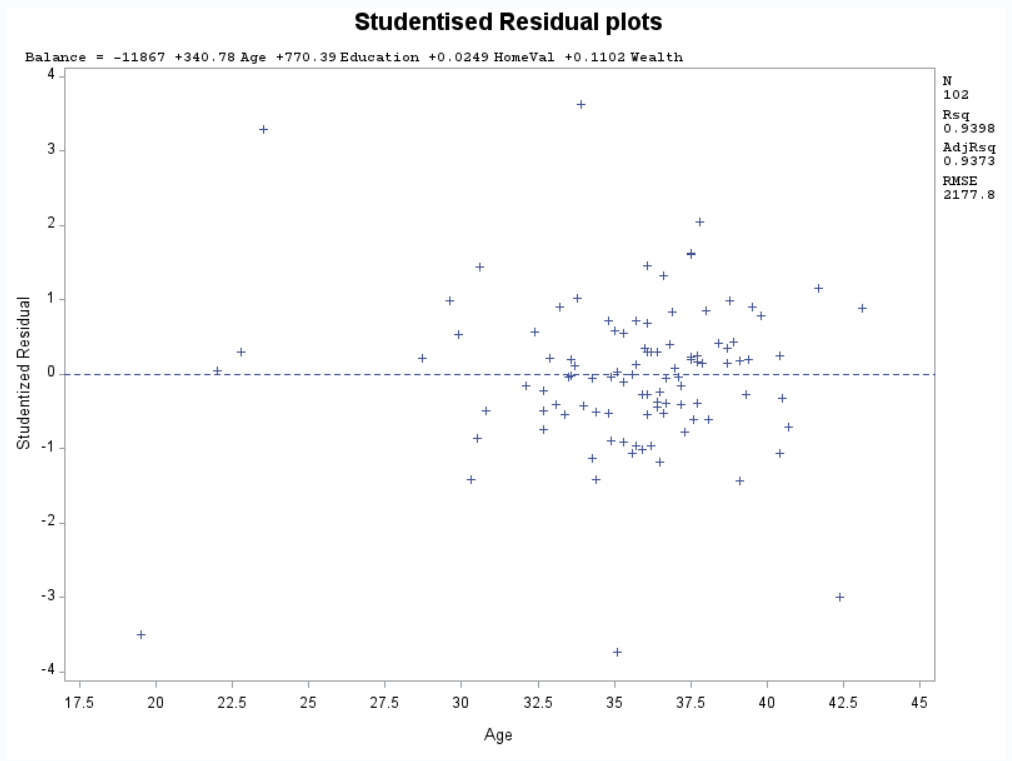
When we compare the R-Squared values and Adjusted R-Squared values for both Models (M1 and M2) we can see that there is a decrease in the values but M2 will be preferred because it doesn’t suffer from multi collinearity.

The Regression Equation for this Model M2 is as follows:

**Balance = -11867 + 340.77\*Age + 770.39\*Education + 0.024\*HomeVal + 0.1103\*Wealth**

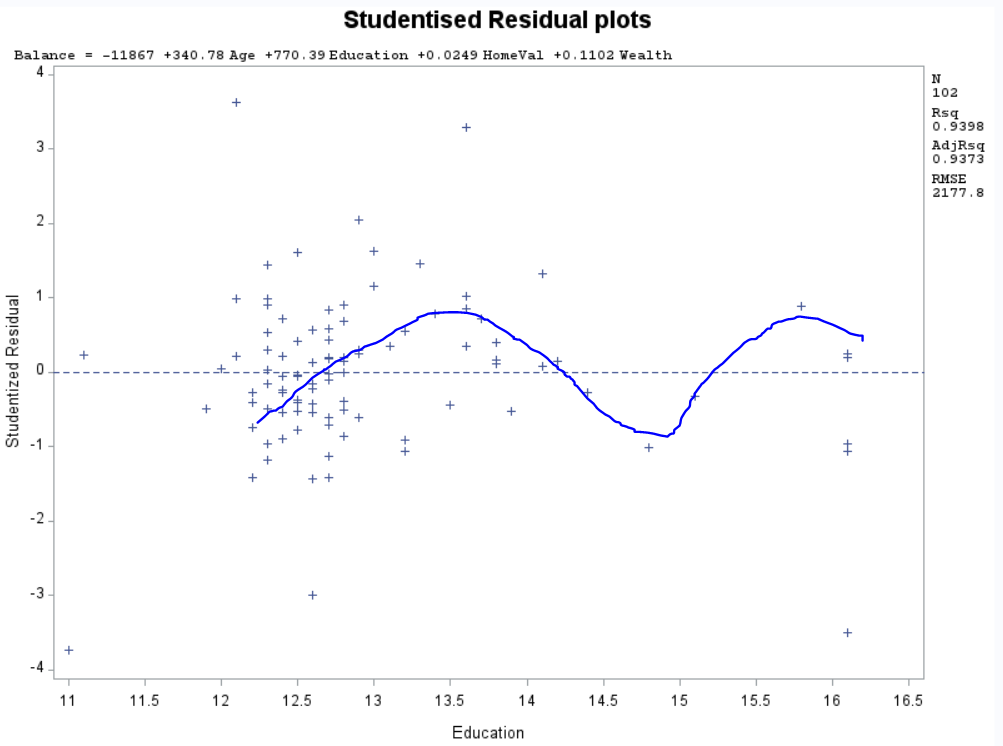
This is my Model M2.

1. The Residual Plots for this model are as follows:



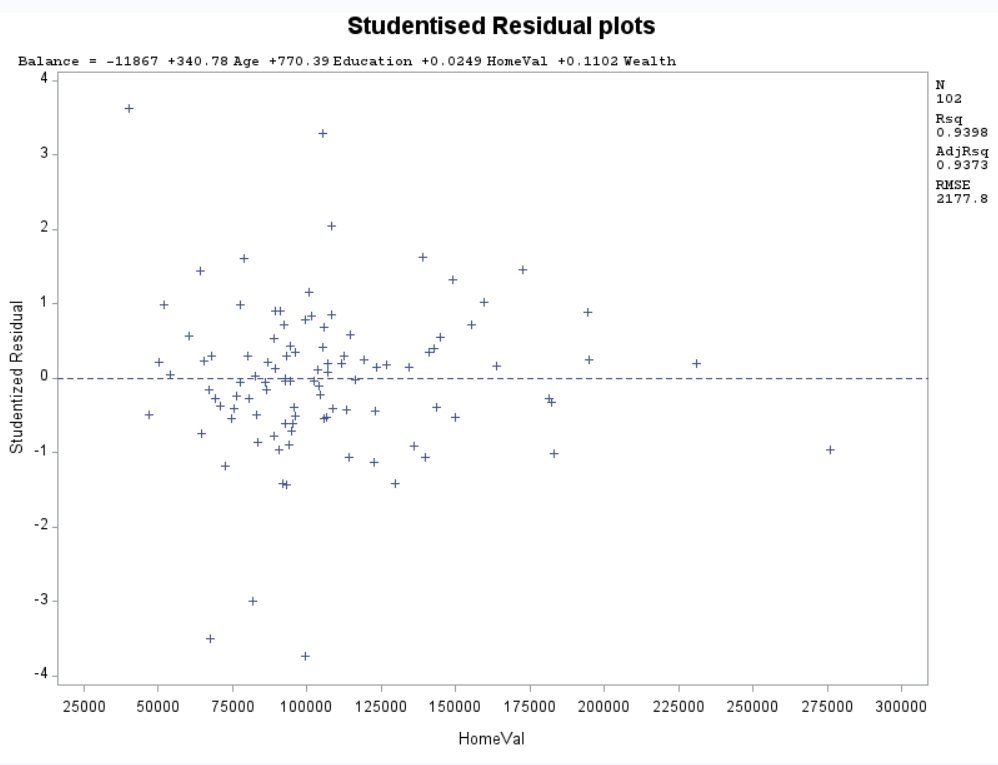
The above is the studentized residual plot for Age. The following observations can be made for the above plot:

* Outliers are the data points which are outside the +/- 3 line. 5 in this case.
* The points seem to funnel outwards when we move from left to right which doesn’t show constant variance.
* There doesn’t seem to be any constant variance which means the points aren’t very independent.
* The majority of points are also concentrated more towards the age group of 30-42.5 and hence doesn’t show much randomness in the model.



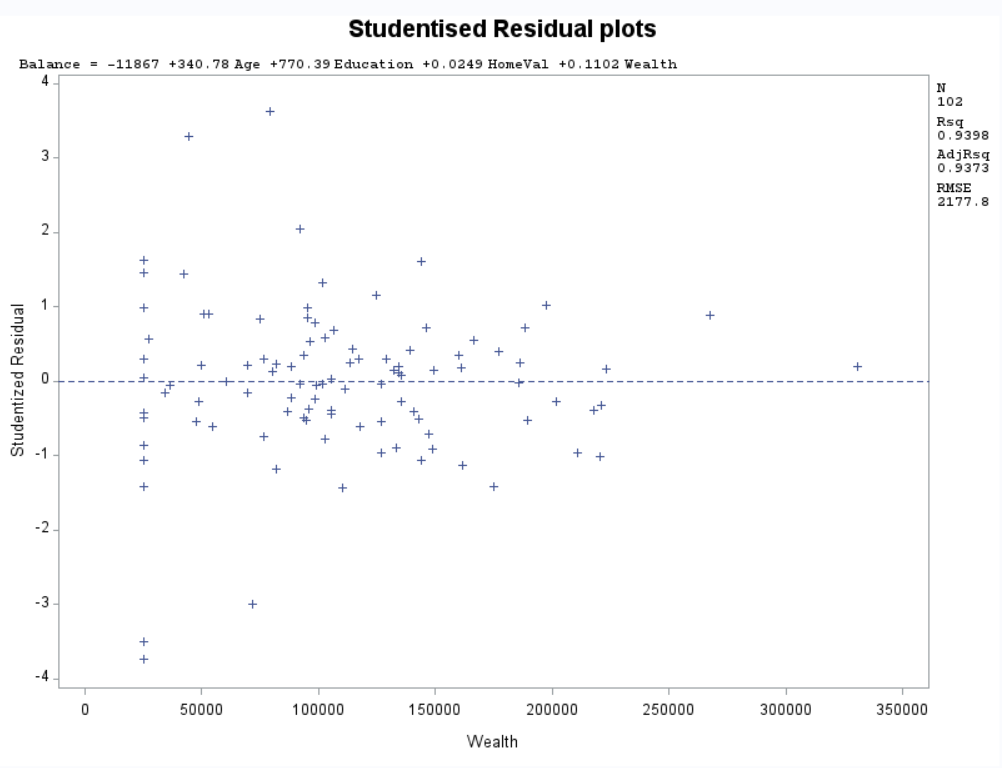
The above is the studentised plot for Education and from this we can make the following observations:

* There are outliers present in this plot. 4 to be exact.
* The scattering of the points is more random when compared to age.
* The plot funnels in as we move from left to right which tells us that there might not be a constant variance.
* The points seem to have lesser independence as there is a kind of pattern present in it which is represented with the blue line.



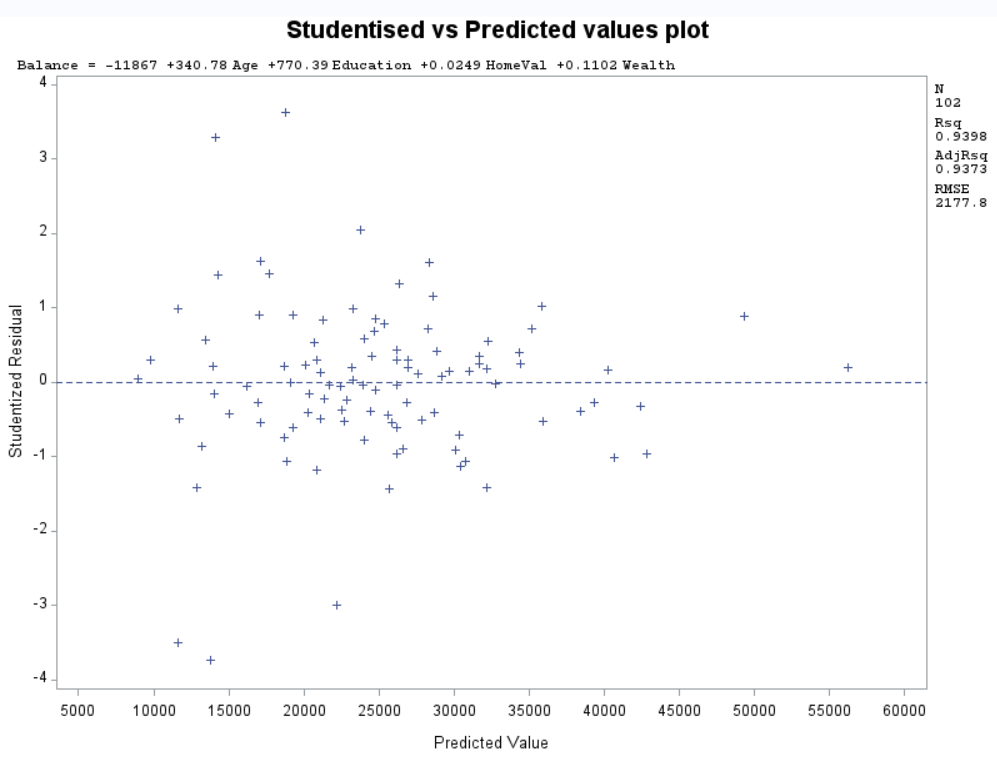
The following observations can be made from the above mentioned Studentised plot for HomeVal:

* Outliers are present.
* The points in HomeVal seem to be scattered more randomly as compared to other predictors and the variance in this is nearly constant. Since there is no pattern or shape that is clearly present we can say that the data points are independent of each other.



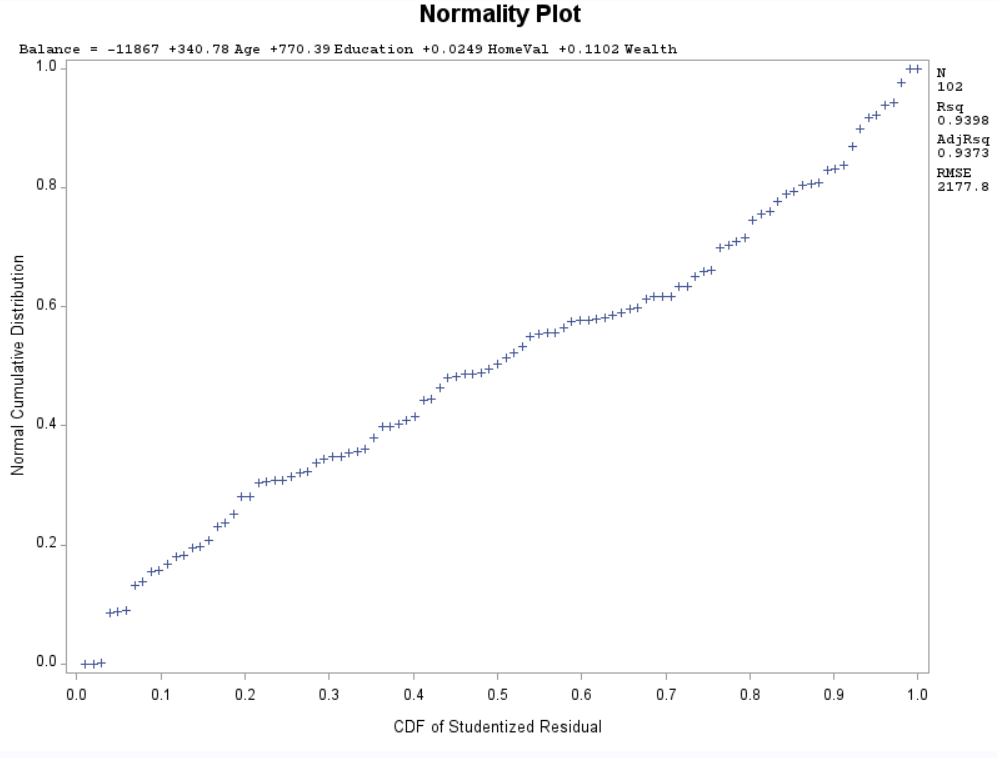
The following observations can be made from the studentised residual plot for Wealth:

* The above plot shows the presence of outliers in the mdoel because there are points above and below the +/-3 line.
* Initially we can see that at around 25000$ wealth mark there are quite a lot of points on a straight line which tells us that there isn’t a constant variance which also tells us that the points aren’t completely independent of each other.



The following points can be inferred from the above mentioned Studentised vs Predicted values plot:

* Outliers are present in this data set. 5 to be exact because there are 4 points which lie above the +/- 3 line and one point which lies on the border of -3 line.
* The points seem to be randomly scattered till 42500 predicted value. Which tells us that there is constant variance and that the points are independent of each other.s



Based on the normality plot we can make the following observations:

* The line is linear with curves at both ends.
* Linearity of the line tells us that the error is distributed normally over all the datapoints.
* The curve at both the ends tells us that the normality distribution of the data points has longer left and right tails which shows the presence of possible outliers in the model.

1. To study the presence of outliers and influential points we run the following code which computes the residual values and the Cook’s D value which can be used to determine whether a point is an outlier or an influential point.

title "Residual and Influential Point analysis";

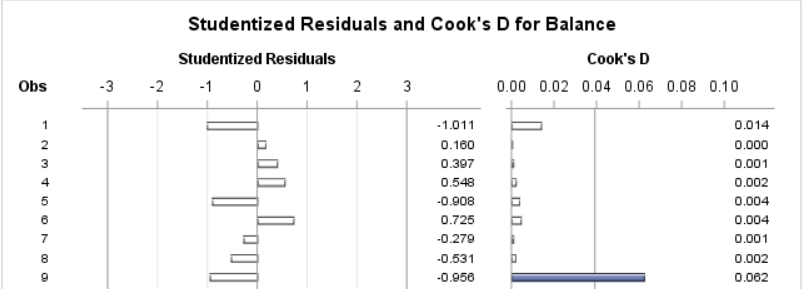
**proc** **reg**;

model balance = age education homeval wealth/ influence r;

**run**;

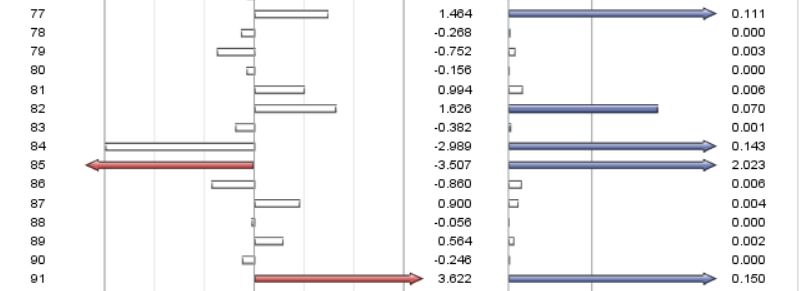
The following output is generated using the above code:

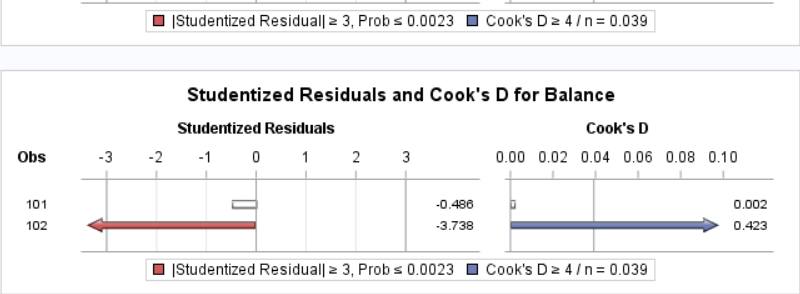
NOTE: The images that follow are of individual data points only.











In the above pictures the points which have red solid color arrow are outliers which lie outside the +/- 3 range in the residual plots.

The points which have blue solid color arrows are influential points. For a point to be considered as an influential point its Cook’s D value must be greater than or equal to 0.039.

The recommendations that I would make is:

* Since the number of outliers isn’t so high I would suggest keeping them rather than removing them.
* Observation number 38, 85 and 91 can be removed because they have been identified as outliers as well as influential points.

1. To compute the standardized coefficients we run the regression model with an added option of stb which is used to compute this value.

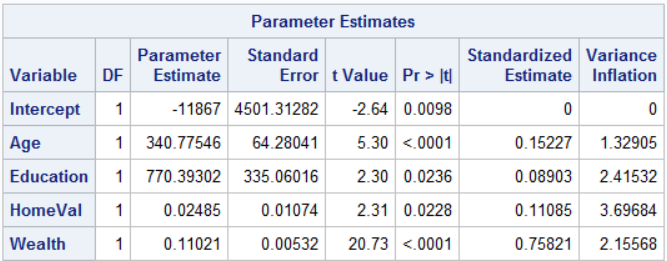
The code for this is as follows:

title "M2 model without income";

**proc** **reg**;

model balance = age education homeval wealth/ vif stb;

**run**;

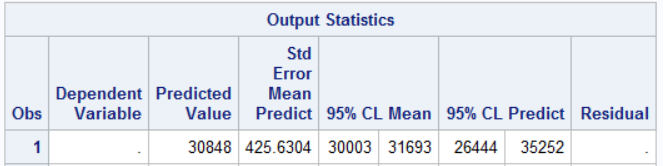


From the above table, we can say that Wealth is the strongest contributor which has the highest influence on balance.

It tells us that with a unit increase in Wealth the balance of the person increases by 0.758 units.

Or if there is an **increase of 1$ in Wealth** then the balance would **increase by 0.11$.**

1. The predicted value is:



The predicted average bank balance is **30848 for the given set of parameters with CI 95%.**

We can say that, with the given set of parameters, **the average bank balance with 95% CI is equal to (30003$, 31693$) and a 95% Prediction Interval equal to (26444$, 35252$).**

The code to generate the predicted value is as follows:

title "Value Prediction using CI = 95%";

**data** pred;

input age education income homeval wealth balance;

datalines;

34 13 64000 140000 160000 .

;

**data** banking;

set pred banking;

**run**;

**proc** **reg**;

model balance = age education homeval wealth/p clm cli alpha = **.05**;

**run**;

The whole code is as follows:

title "Assignment - 4 - Banking Dataset Complete";

**proc** **import** datafile ="Bankingfull.txt" out = banking replace;

delimiter = '09'x;

getnames = yes;

datarow = **2**;

**run**;

title "Scatter plots between the 5 variables";

**proc** **sgscatter**;

MATRIX balance age education income homeval wealth;

**run**;

title "Individual plots";

**proc** **plot**;

plot balance\*(age education income homeval wealth);

**run**;

title "Pearson's Correlation Values";

**proc** **corr**;

var balance age education income homeval wealth;

**run**;

title "M1 model which has all the 5 variables and vif values";

**proc** **reg**;

model balance = age education income homeval wealth/ vif stb;

**run**;

title "M2 model without income";

**proc** **reg**;

model balance = age education homeval wealth/ vif stb;

**run**;

title "M3 model without wewalth";

**proc** **reg**;

model balance = age education income homeval/ vif stb;

**run**;

title "M4 model without income";

**proc** **reg**;

model balance = age wealth/ vif stb;

**run**;

title "M5 model without income";

**proc** **reg**;

model balance = age income/ vif stb;

**run**;

title "M6 model without income";

**proc** **reg**;

model balance = age income wealth/ vif stb;

**run**;

title " Studentised Residual plots";

**proc** **reg**;

model balance = age education homeval wealth;

plot student.\*(age education homeval wealth);

**run**;

title "Studentised vs Predicted values plot";

**proc** **reg**;

model balance = age education homeval wealth;

plot student.\*predicted.;

**run**;

title "Normality Plot";

**proc** **reg**;

model balance = age education homeval wealth;

plot npp.\*student.;

**run**;

title "Residual and Influential Point analysis";

**proc** **reg**;

model balance = age education homeval wealth/ influence r;

**run**;

title "Value Prediction using CI = 95%";

**data** pred;

input age education income homeval wealth balance;

datalines;

34 13 64000 140000 160000 .

;

**data** banking;

set pred banking;

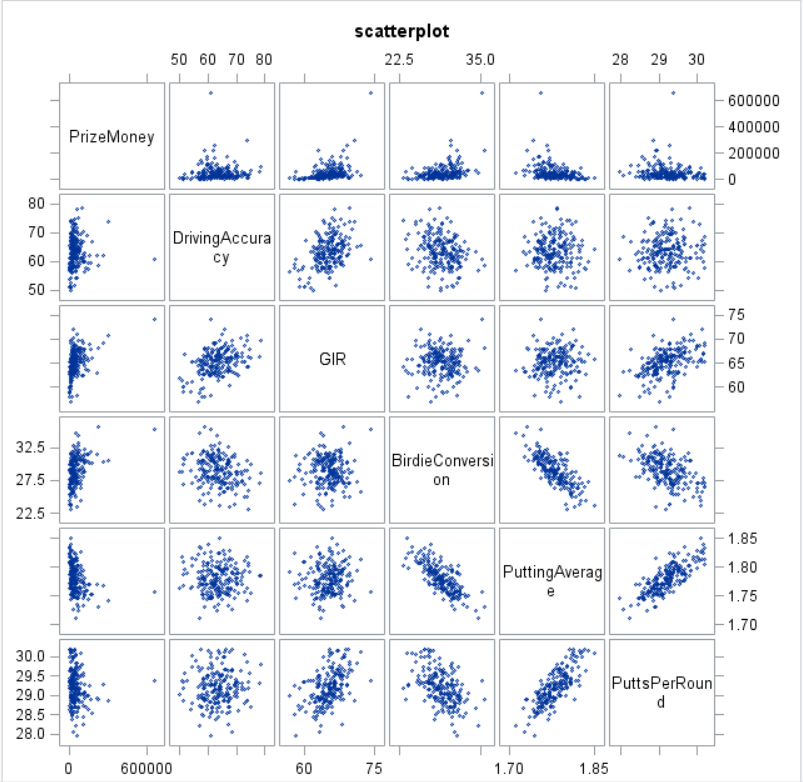
**run**;

**proc** **reg**;

model balance = age education homeval wealth/p clm cli alpha = **.05**;

**run**;

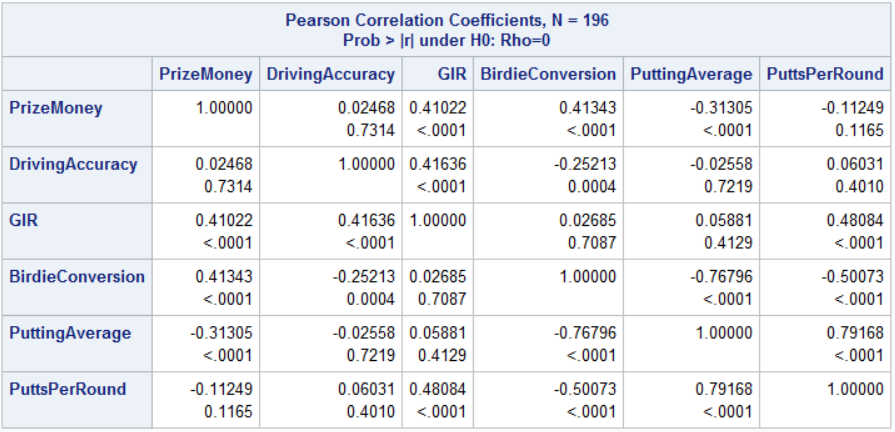
1. Problem 2 – PGA Tour
   1. Scatterplots Matrix



PrizeMoney **doesn’t seem to have any Linear relation** with any of the other Independent variables.

To further provide evidence for this we will also calculate correlation values just to observe if the independent variables are correlated with the PrizeMoney (Dependent Varaible) or not.

The following are the correlation values for the variables:

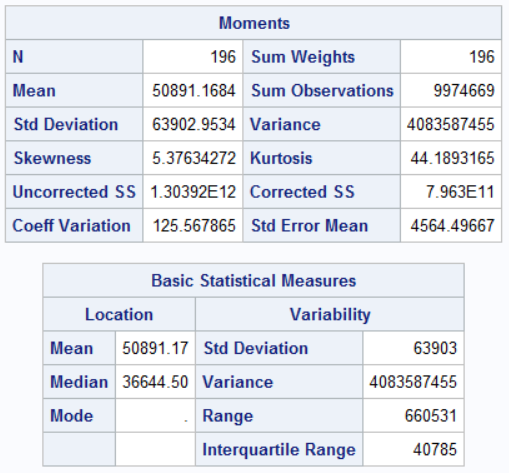


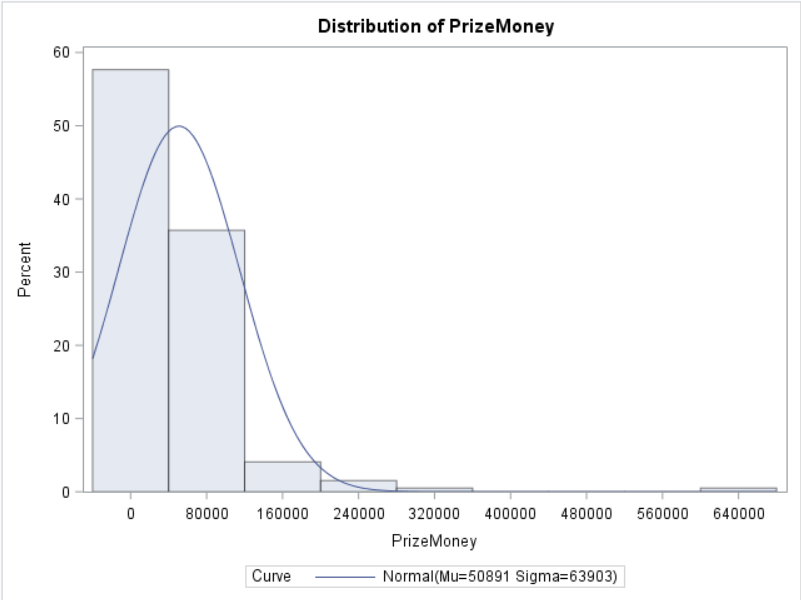
From the above values, we can see that PrizeMoney **isn’t strongly correlated** to any other variables.

PrizeMoney is **positively correlated** to DrivingAccuracy, GIR and BirdieConversion but is **negatively correlated** to PuttingAverage and PuttsPerRound.

Overall, we can say that some kind of Transformation has to be done in order to come up with a meaningful model.

* 1. The Distribution of PrizeMoney:





The following observations can be made from the above histogram and the basic statistical measures:

* The first thing that we can observe from the histogram is that the distribution is **positively skewed or Right Skewed**. The **Mean > Median** which also supports this claim.
* IQR = 40785 and hence anything which lies beyond **Q3+1.5\*IQR (58006.5 + 1.5\*40875) = 119,319$** will be an outlier. Based on the histogram we can see that there are outliers present in this distribution.
* The kurtosis value is greater than 3 meaning the tails are lighter and the peak is high.
* From the moments table, we can see that the value of skewness is **greater than 5** which means the **Right Tail is longer than the left tail** and that the distribution is **positively skewed**.
  1. To perform the log transformation on PrizeMoney we run the following code:

**data** pga;

set pga;

ln\_prize = log(PrizeMoney);

**run**;

Now to create the Histogram analysis we run the following piece of code:

title "Analysis of log(PrizeMoney)";

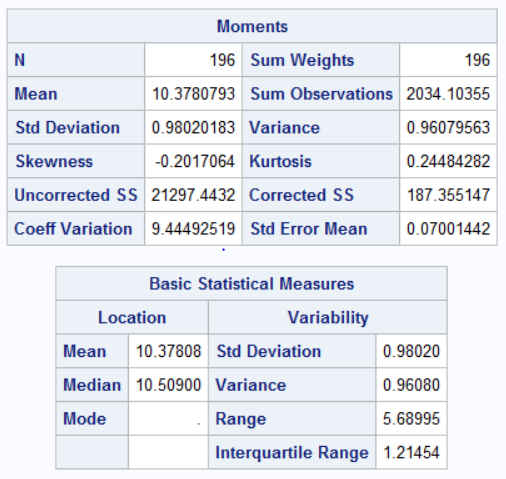
**proc** **univariate** normal;

var ln\_prize;

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

**run**;

The following output is generated by SAS:



Since the **Mean < Median**, we can say that the distribution is **slightly skewed towards the left or Negatively Skewed**. This can also be observed by looking at the Moments table in which the value of skewness is Negative.

The **Kurtosis value is < 3 (0.244)** which means that the **tails are more spread out** which can be observed from the histogram.

The code to generate the Histogram is as follows:

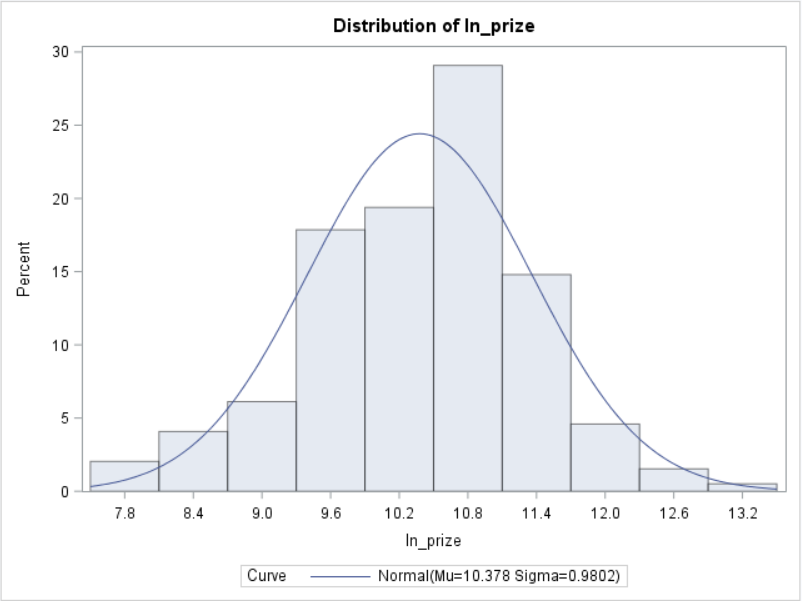
title "Analysis of log(PrizeMoney)";

**proc** **univariate** normal;

var ln\_prize;

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

**run**;



The IQR for this is 1.21454, which means that any data points outside the range of (Q1-1.5\*IQR, Q3+1.5\*IQR) = **(7.932, 12.789)** would be considered as an outlier.

Based on the above Range of IQR we can say that this **distribution contains outliers**.

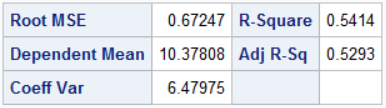
* 1. The Regression Model:
     1. The first model which I created will have all the variables and the code to generate this model is as follows:

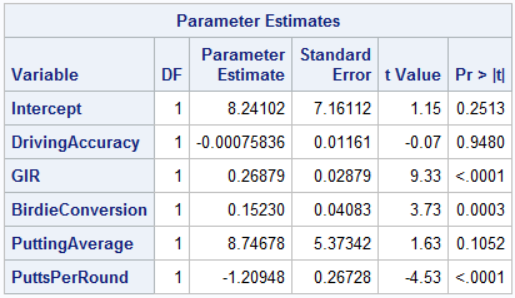
title "Regression Model including all variables";

**proc** **reg**;

model ln\_prize = Drivingaccuracy gir birdieconversion puttingaverage puttsperround;

**run**;





As we can see from the above table that DrivingAccuracy and PuttingAverage have p-value greater than 0.05 and hence we wouldn’t be choosing this model.

* + 1. This model excludes only Driving Accuracy:

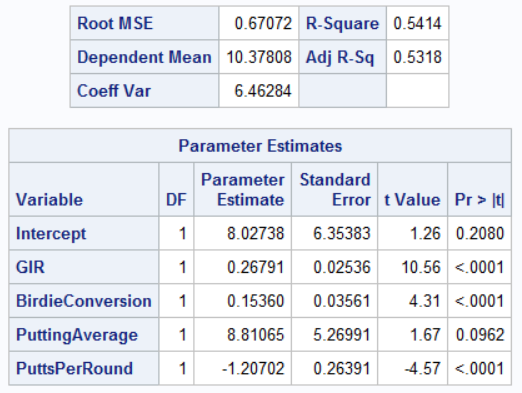
The code to generate this model in SAS is as follows:

title "Regression Model excluding DrivingAccuracy";

**proc** **reg**;

model ln\_prize = gir birdieconversion puttingaverage puttsperround;

**run**;



From the above table we can see that PuttingAverage has p-value which is still greater than 0.05 and hence we will discard this model also as it contains insignificant variable.

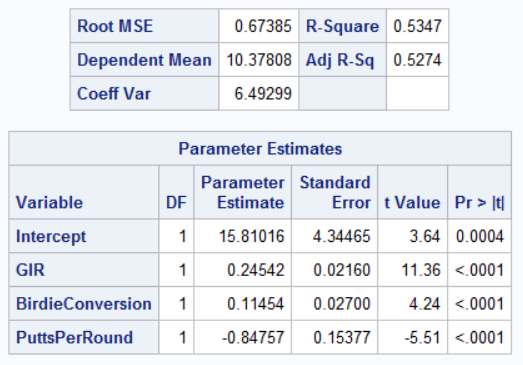
* + 1. In this model I’ve also excluded PuttingAverage. The code to generate this model is as follows:

title "Regression Model excluding DrivingAccuracy and puttingaverage";

**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround;

**run**;



From the above table, we can say that this model no longer contains any insignificant variables and hence we will fix this model as our final model and perform further analysis using this model.

To generate the residual plot, we run the following piece of code:

title "Studentised Residual Analysis";

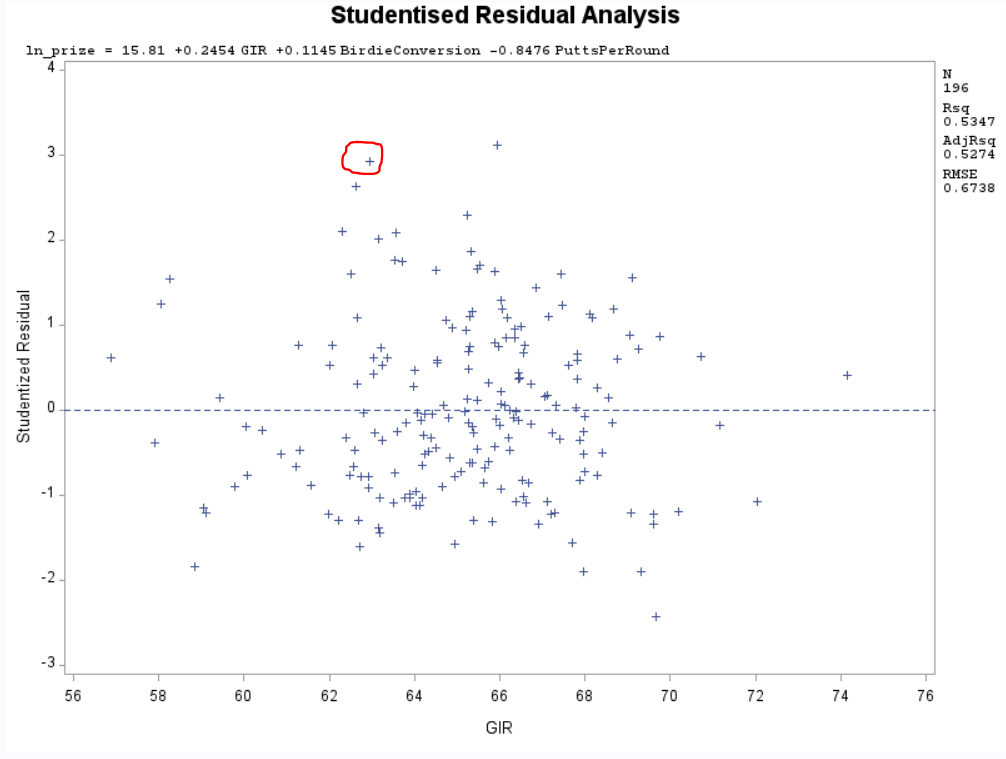
**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround;

plot student.\*(gir birdieconversion puttsperround);

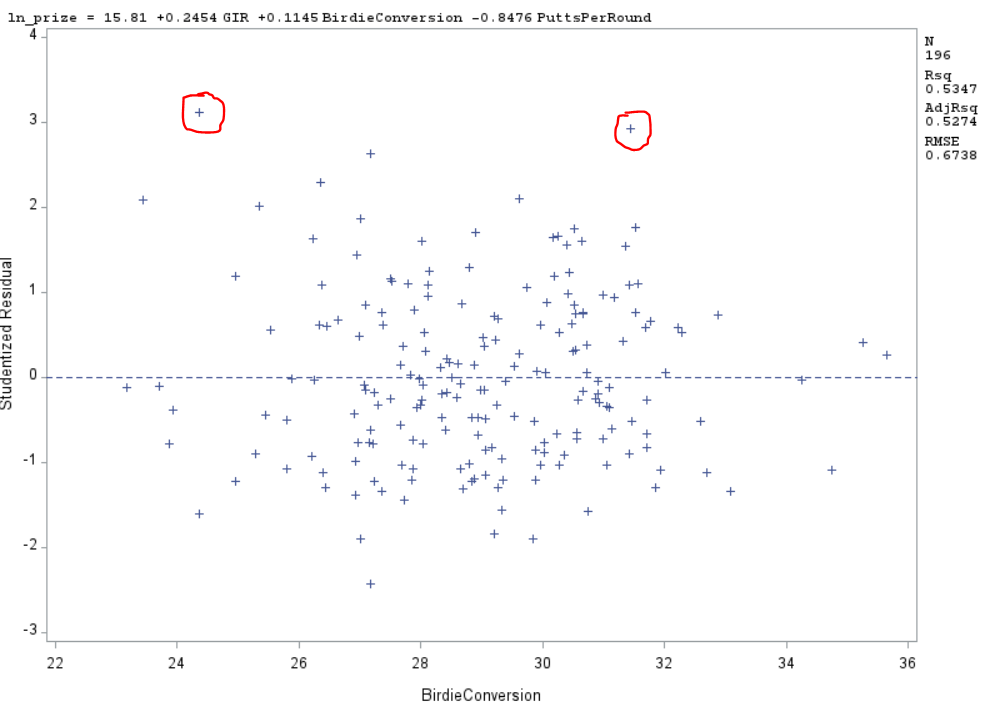
**run**;

We get the following plots as the output by running the above code:



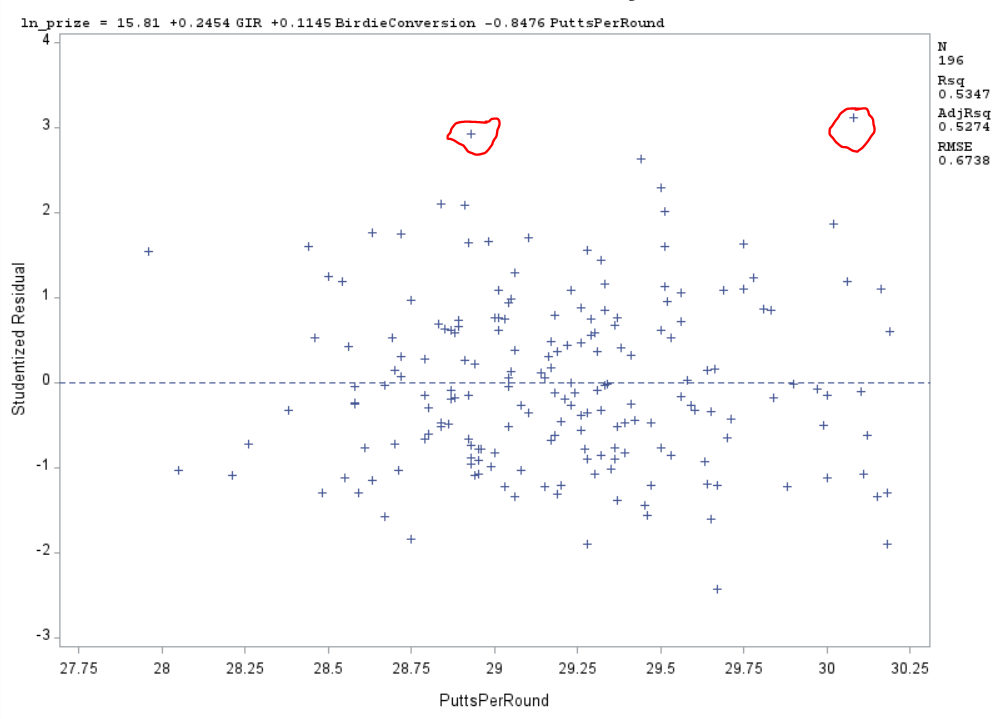
From the above Residual Plot for GIR we can make the following observations:

* The point which has been circled lies on the boundary and hence it may or may not be treated as an outlier. There is a point which is just above the +3 line and that is treated as an outlier.
* The whole plot seems nearly random and there seems to be somewhat constant variance and the points look to be independent of each other.
* Overall there doesn’t seem to be any kind of pattern that is significant in the above figure.



From the above Studentised Residual plot of Birdie Conversion we can make the following observations:

* The two circled points lie very close to the boundary line of +3 in the residual plot. The one from the left is above the +3 line and can be treated as an outlier and the one in the right may or may not be treated as an outlier as it lies very close to the boundary.
* The points seem to be scattered randomly and there seems to be somewhat constant variance and the points look to be independent of each other.
* There is no pattern which is being significant in the plot and hence we can say that the points are independent of each other.



From the Studentised plot of PuttsPerRound we can make the following observations:

* The two circled points feel like outliers as they lie on the boundary and hence closer observations of these points is required.
* The points seem to be randomly scattered and there isn’t a pattern which is significant in the plot.
* The points seem to have constant variance and look independent of each other.

To generate the Studentised Residual Plot vs the Predicted value we run the following piece of code in SAS:

title "Residual Plot (Predicted Value)";

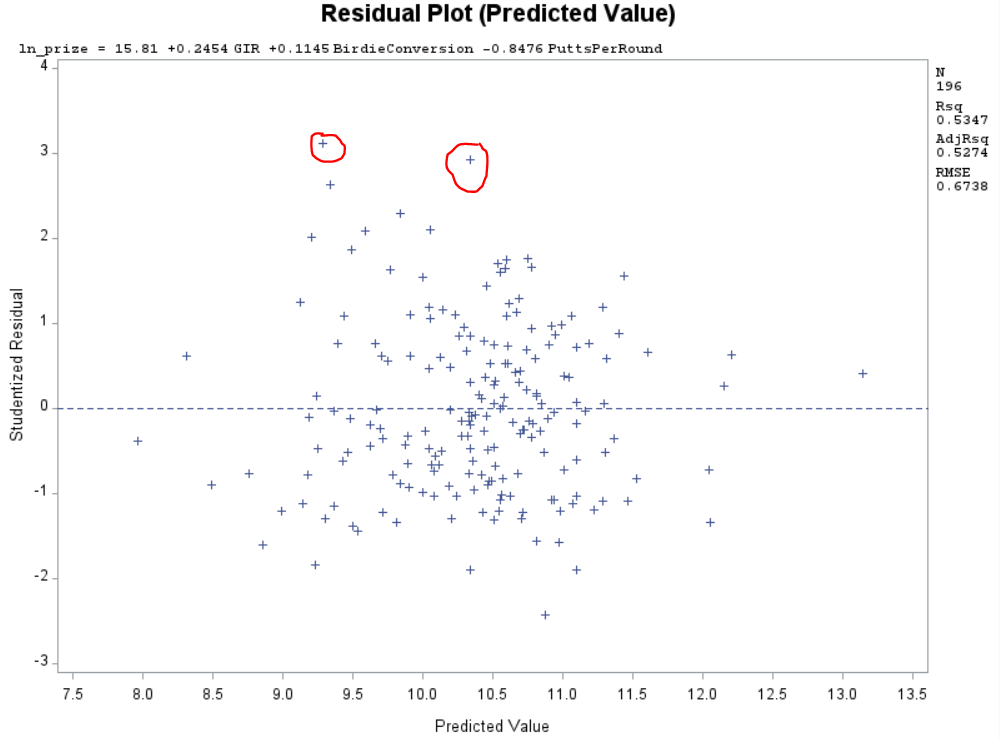
**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround;

plot student.\*predicted.;

**run**;

The above code generates the following output:



From the above plot, we can make the following observations:

* The points seem to funnel outwards in the beginning but apart from that the points seem to be randomly scattered and there isn’t a pattern which is significant in the whole distribution.
* The two circled points lie on the boundary and hence may or may not be considered as outliers and closer study of this points is required.
* There seem to be almost a constant variance between the points and the points look to be independent of each other.

To generate the normality plot we run the following piece of code in SAS:

title "Normality Plot";

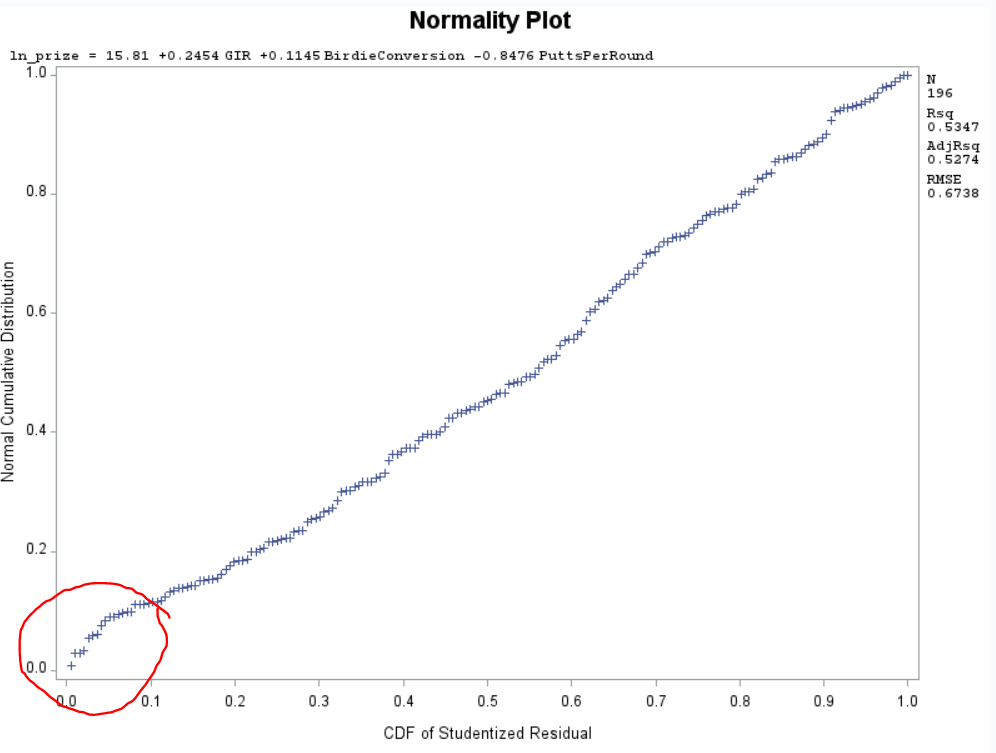
**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround;

plot npp.\*student.;

**run**;

The following output is generated when we run the above code:



From the above Normality plot we can make the following observations:

* The slight curve at the bottom gives evidence to the presence of outliers in the dataset which was also observed in the various residual plots.
* The normality curve is more or less Linear and hence we can say that the error is normally distributed over all the predictions.

Overall the plots have validated the analysis we have made before which is associated with this dataset.

To analyze if there are any outliers or influential points present in the dataset we run the following piece of code in SAS:

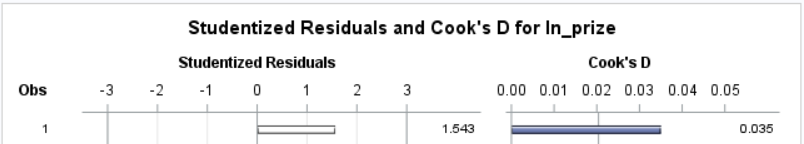
title "Residual and Influential Point analysis";

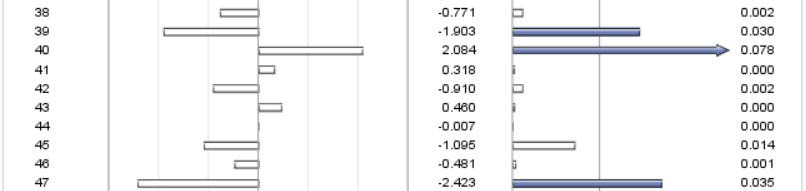
**proc** **reg**;

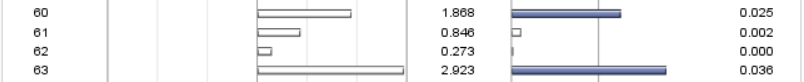
model ln\_prize = gir birdieconversion puttsperround / influence r;

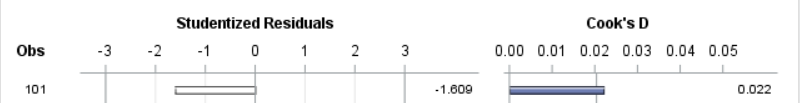
**run**;

The above generates the following output:

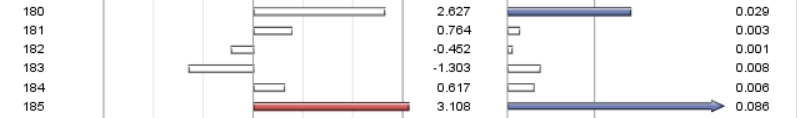














From the above plot, we can see that any data point which has Cook’s D value greater than 0.02 will be treated as an influential point and any data point which has Studentised Residual value greater than +/- 3 will be treated as an outlier.

From the above-mentioned residual plots and Cook’s D value plot we can say that the **data point 185** is an outlier as well as an **influential point** and can be removed from the dataset because it’s only a single point.

All other points which have Solid Blue Arrows/Rectangles are considered as influential points. The points which need to be **investigated are points: 1, 39, 40, 47, 63, 180 and 185** because the **Cook’s D value associated with these points is above 0.02** and hence in **future these points may become outliers** as the Studentized Residual values for these points is also close to 3 and by adding more records to the dataset or by adding more variables these points behaviors may change and hence it is advisable to keep a check on them.

The Final Model Equation for this model is as follows:

**Ln\_prize = 15.81 + 0.2454\*GIR + 0.1145\*BirdieConversion – 0.8476\*PuttsPerRound**

The above model is the best fit because of the following reasons:

* Of all the models that I considered, this model had no insignificant contributors and the Adjusted R-Squared Value was also close to the other models which had included insignificant attributes.
* This model had just one outlier and was successful in fitting all the other data points.
* The F- Value associated with this model was also the highest when compared to the other models. The F-value for the **model -1 was 44.86**, for **Model -2 F-Value was 56.37** and for the **Model -3** which I had chosen the **F-value is 73.54 and the p-value for all the models was <0.0001.**
  1. From the above equation, we can see that the **Beta Coefficient for GIR is 0.2454** so a 1% increase in GIR would result in the **Rise of PrizeMoney by** ((e^0.2454) – 1)\*100 => (1.2781-1)\*100 => **27.81%.**

Therefor, **1% increase in GIR would result in 27.81% increase in PrizeMoney.**

* 1. To predict we run the following code:

title "Value Prediction using CI = 95%";

**data** pred;

input prizemoney Drivingaccuracy gir birdieconversion puttingaverage puttsperround;

datalines;

. 64 67 28 1.77 29.16

;

**data** pga;

set pred pga;

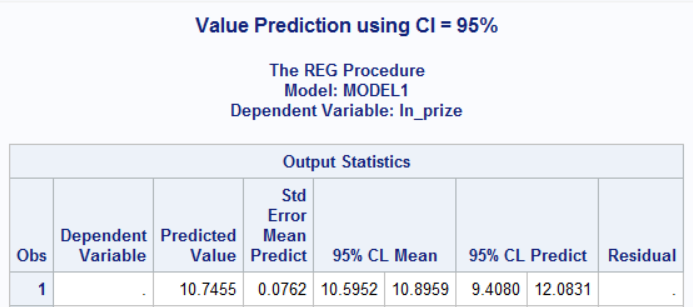
**run**;

**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround/p clm cli alpha = **.05**;

**run**;

We can see that the predicted row has been added to the top of the pga dataset.



From the above we can see that **ln\_prize = 10.7455 and the PrizeMoney will be e^10.7455 which is 46420.66$.**

Thus, we can say that with the given set of parameters that average **ln\_prize will be in the range of (10.5952,10.8959) with 95% CI and with 95% prediction interval, ln\_prize will lie in the range of (9.4080,12.0831).**

Thus, we can say that with the given set of parameters that average **PrizeMoney will be in the range of (39942$, 53954.69$) with 95% CI and with 95% prediction interval, PrizeMoney will lie in the range of (12185.47$, 176857.57$).**

* 1. The whole code for this dataset is as follows:

title "PGA Tour";

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

delimiter=',';

getnames = yes;

**proc** **print**;

**run**;

title "scatterplot";

**proc** **sgscatter**;

MATRIX prizemoney Drivingaccuracy gir birdieconversion puttingaverage puttsperround;

**run**;

title "Correlation values";

**proc** **corr**;

var prizemoney Drivingaccuracy gir birdieconversion puttingaverage puttsperround;

**run**;

title "Analysis of PrizeMoney";

**proc** **univariate** normal;

var PrizeMoney;

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

**run**;

title "Log conversion";

**data** pga;

set pga;

ln\_prize = log(PrizeMoney);

**run**;

title "Analysis of log(PrizeMoney)";

**proc** **univariate** normal;

var ln\_prize;

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

**run**;

title "Regression Model including all variables";

**proc** **reg**;

model ln\_prize = Drivingaccuracy gir birdieconversion puttingaverage puttsperround;

**run**;

title "Regression Model excluding DrivingAccuracy";

**proc** **reg**;

model ln\_prize = gir birdieconversion puttingaverage puttsperround;

**run**;

title "Regression Model excluding DrivingAccuracy and puttingaverage";

**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround;

**run**;

title "Regression Model excluding DrivingAccuracy and puttingaverage";

**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround/vif stb;

**run**;

title "Studentised Residual Analysis";

**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround;

plot student.\*(gir birdieconversion puttsperround);

**run**;

title "Residual Plot (Predicted Value)";

**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround;

plot student.\*predicted.;

**run**;

title "Normality Plot";

**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround;

plot npp.\*student.;

**run**;

title "Residual and Influential Point analysis";

**proc** **reg**;

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

**run**;

title "Value Prediction using CI = 95%";

**data** pred;

input prizemoney Drivingaccuracy gir birdieconversion puttingaverage puttsperround;

datalines;

. 64 67 28 1.77 29.16

;

**data** pga;

set pred pga;

**run**;

**proc** **reg**;

model ln\_prize = gir birdieconversion puttsperround/p clm cli alpha= **.05**;

**run**;