Ramadan Gannud
Assignment-4

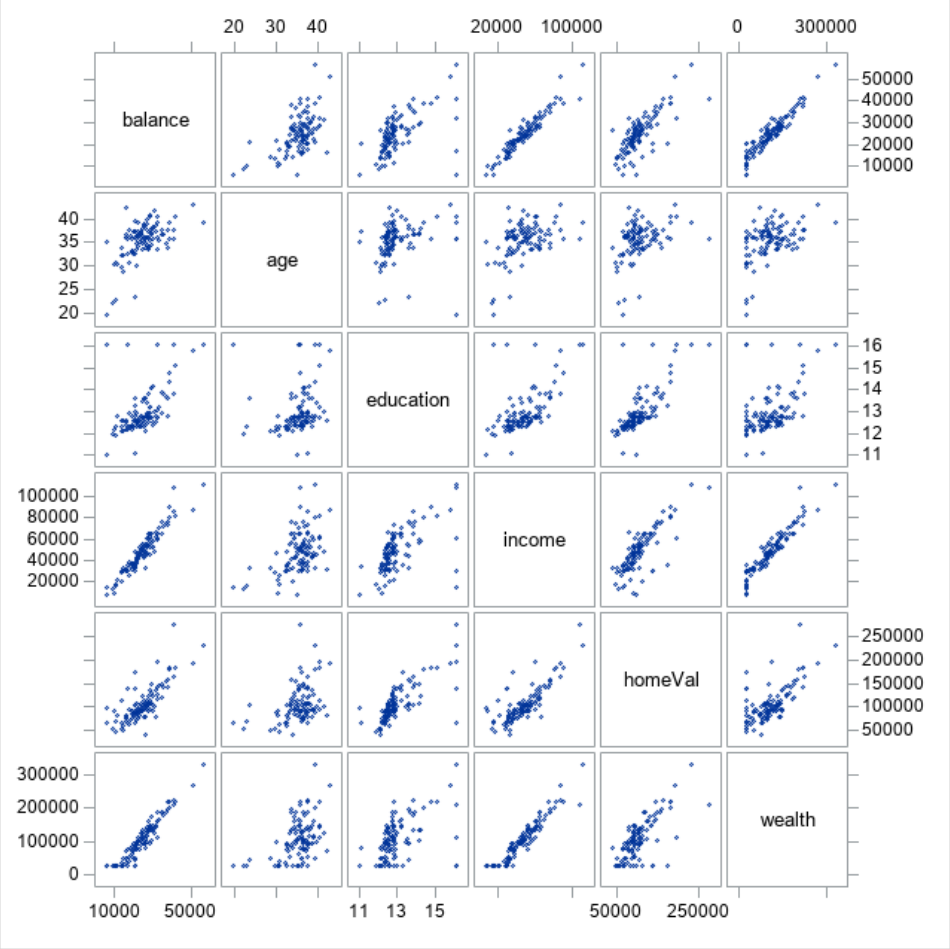
PROBLEM 1 [16 pts] – to be answered by everyone

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.

a) 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)



In the balance vs income and balance vs wealth, plots are creating a high positive correlation and the association appear to be linear. In the other scatterplot where balance was plotted against homeVal, the scatterplot looks linear with low positive correlation. In the scatterplot where balance was plotted against education, there is a very low positive correlation with so many outliers. Also, there is a very low positive correlation between balance and age as well.

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.

Pearson Correlation Coefficients, N = 102
Prob > |r| under H0: Rho=0

	balance	age	education	income	homeVal	wealth
balance	1.00000	0.56547 <.0001	0.55488 <.0001	0.95168 <.0001	0.76639 <.0001	0.94871 <.0001
age	0.56547 <.0001	1.00000	0.17341 0.0813	0.47715 <.0001	0.38649 <.0001	0.46809 <.0001
education	0.55488 <.0001	0.17341 0.0813	1.00000	0.57539 <.0001	0.75352 <.0001	0.46941 <.0001
income	0.95168 <.0001	0.47715 <.0001	0.57539 <.0001	1.00000	0.79536 <.0001	0.94667 <.0001
homeVal	0.76639 <.0001	0.38649 <.0001	0.75352 <.0001	0.79536 <.0001	1.00000	0.69848 <.0001
wealth	0.94871 <.0001	0.46809 <.0001	0.46941 <.0001	0.94667 <.0001	0.69848 <.0001	1.00000

There is a significant linear association between account balance and income since the correlation value is 0.95 which is high. There is also a strong linear association between the account balance variable and wealth because correlation value is high as well 0.94. homeVal show a lower positive correlation 0.76 which shows a smaller linear association. Age and education have close correlation values which are 0.56 and 0.55 respectively. That does not show a significant linear relationship between the Y variable “balance” and both of X values “age and education”.

c) 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.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-10711	4260.97631	-2.51	0.0136	.	0
age	1	318.66496	60.98611	5.23	<.0001	0.74473	1.34276
education	1	621.86035	318.95952	1.95	0.0541	0.40705	2.45671
income	1	0.14632	0.04078	3.59	0.0005	0.06711	14.90172
homeVal	1	0.00918	0.01104	0.83	0.4075	0.22815	4.38300
wealth	1	0.07433	0.01119	6.64	<.0001	0.09333	10.71428

Balance = $\beta_0 + \beta_1$*age + β_2*education + β_3*income + β_4*homeVal + β_5*wealth + e

Balance = -10711 + 318.66*age + 621.86*education + 0.14*income + 0.0091*homeVal +0.07433*wealth + e

Diagnosing Multicollinearity:

1. Scatterplot matrix and Pearson correlation matrix for each pair of x variables:
   There is a serious collinearity problem between income and wealth because the correlation value between them is 0.94.
2. Compute Variance Inflation Factor (VIF):
   VIF for income = 14.9 > 10
   VIF for wealth = 10.7 > 10
   These VIFs suggest collinearity for income and wealth.
3. Compute Tolerance value (TOL):
   TOL for income = 0.067 < 0.1
   TOL for wealth = 0.93 < 0.1
   These TOLs suggest collinearity for income and wealth.

In this case, I would exclude income variable since it has the highest Variance Inflation Factor.

Obs	age	education	homeVal	wealth	balance
1	35.9	14.8	183104	220741	38517
2	37.7	13.8	163843	223152	40618
3	36.8	13.8	142732	176926	35206

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:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7180778778	1795194694	378.50	<.0001
Error	97	460065367	4742942		
Corrected Total	101	7640844145			

Root MSE	2177.82964	R-Square	0.9398
Dependent Mean	24888	Adj R-Sq	0.9373
Coeff Var	8.75056		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-11867	4501.31282	-2.64	0.0098	.	0
age	1	340.77546	64.28041	5.30	<.0001	0.75241	1.32905
education	1	770.39302	335.06016	2.30	0.0236	0.41402	2.41532
homeVal	1	0.02485	0.01074	2.31	0.0228	0.27050	3.69684
wealth	1	0.11021	0.00532	20.73	<.0001	0.46389	2.15568

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7235179873	1447035975	342.44	<.0001
Error	96	405664272	4225669		
Corrected Total	101	7640844145			

Root MSE	2055.64333	R-Square	0.9469
Dependent Mean	24888	Adj R-Sq	0.9441
Coeff Var	8.25962		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-10711	4260.97631	-2.51	0.0136	.	0
age	1	318.66496	60.98611	5.23	<.0001	0.74473	1.34276
education	1	621.86035	318.95952	1.95	0.0541	0.40705	2.45671
income	1	0.14632	0.04078	3.59	0.0005	0.06711	14.90172
homeVal	1	0.00918	0.01104	0.83	0.4075	0.22815	4.38300
wealth	1	0.07433	0.01119	6.64	<.0001	0.09333	10.71428

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

adj-R^2 for M1 = 0.9441

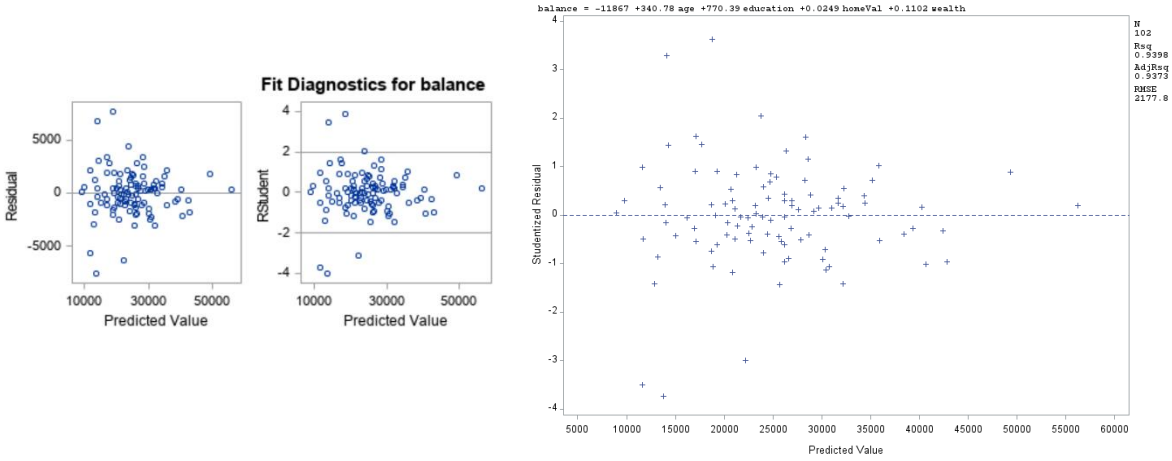
adj-R^2 for M2 = 0.9373

Model 1 has the largest adj-R2. A higher Adj-R^2 indicates a good model.

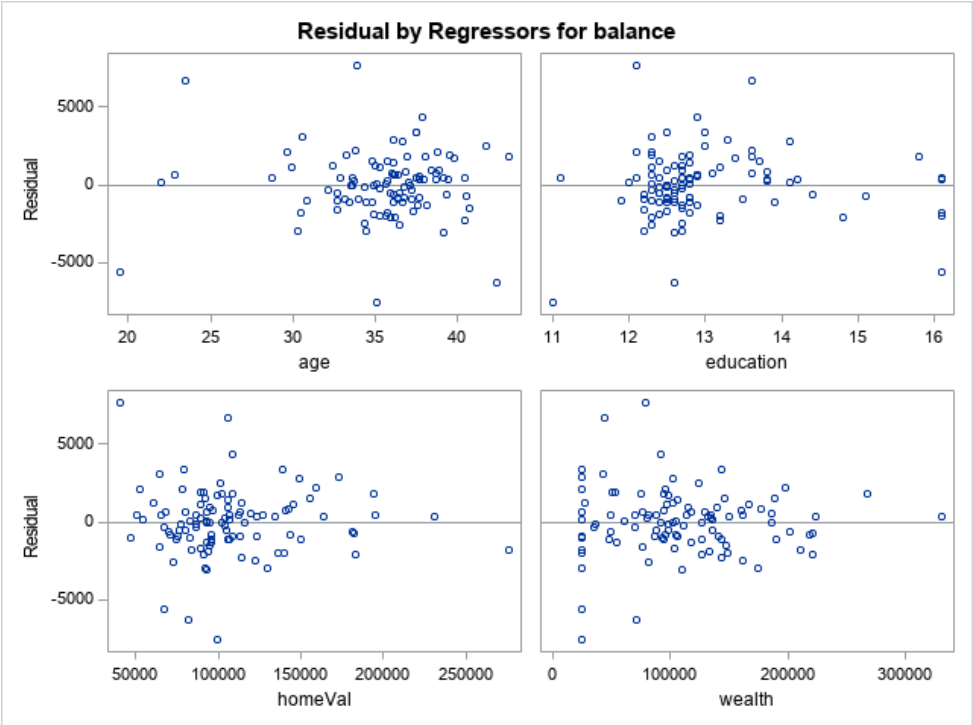
b. Create residual plots for M2 (standardized residuals vs predicted; standardized residuals vs x-variables; and normal plot of residuals). Analyze the residual plots to check if the regression model assumptions are met by the data. Include the relevant output and discuss your analysis.

Assumptions of Constant Variance and Independence:

1. Plot residuals vs predicted values.



2. Plot residuals vs each x-variable.

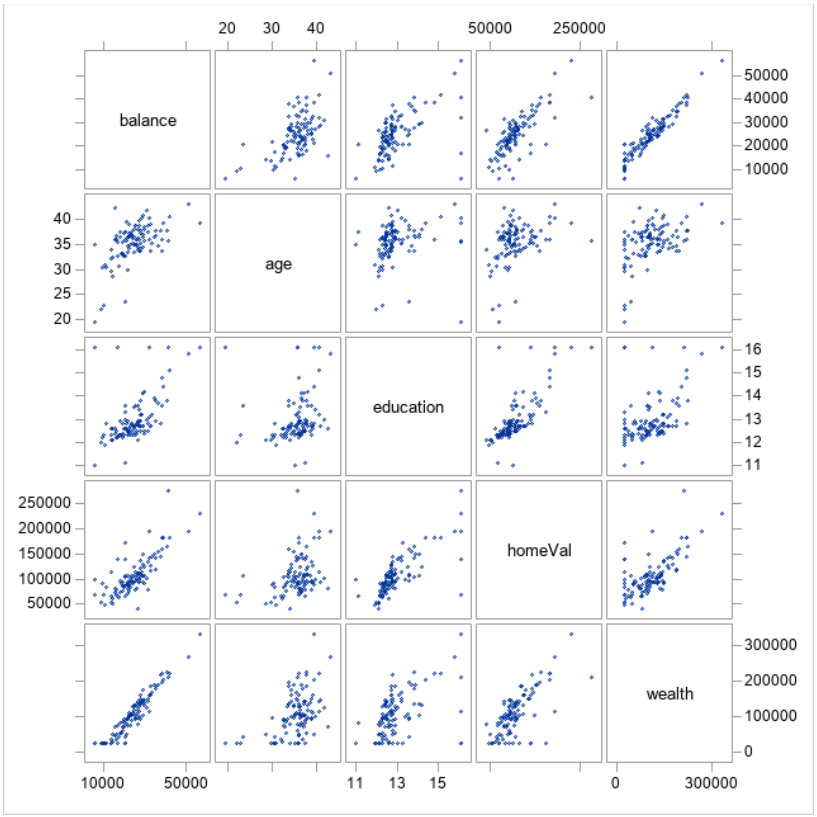


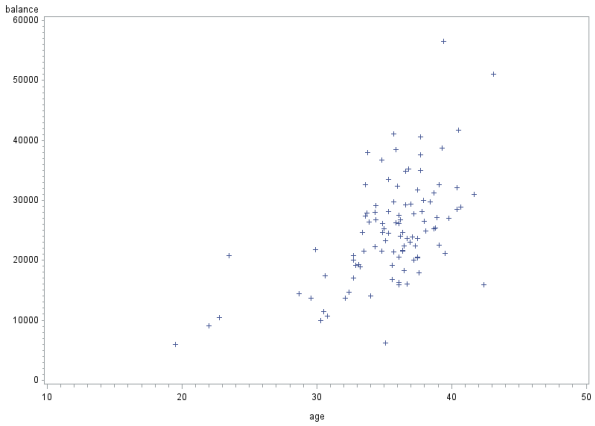
Points are randomly scattered, and residual analysis show no concern for the model fit.

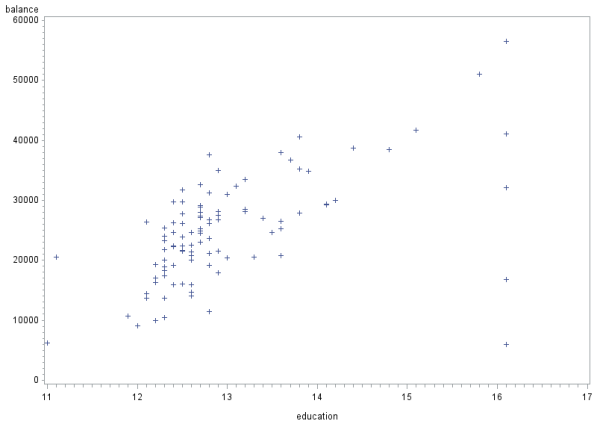
Assumptions for linearity:

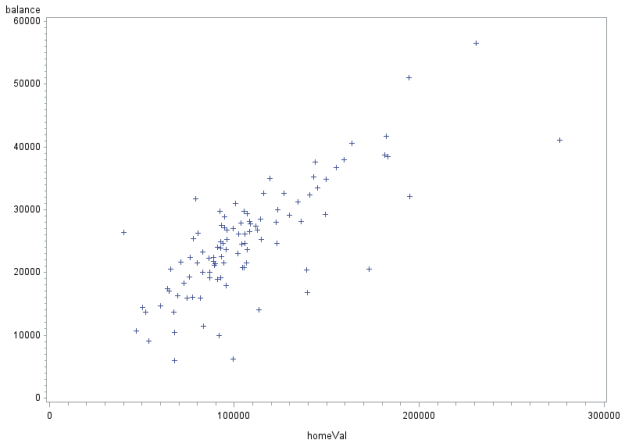
1. Scatterplot for each x-variable.

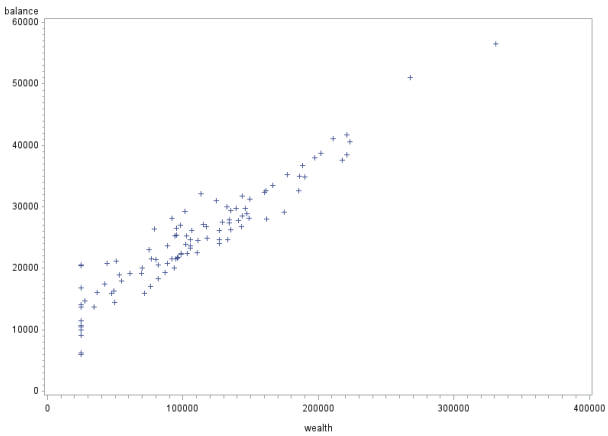
Only the wealth scatterplot shows some linearity and the association appear to be linear as it's shown in the next graphs. homeVal shows a low positive linear correlation.


balance
age
education
homeVal
wealth
20 30 40
50000 250000
50000 40000 30000 20000 10000
40 35 30 25 20
16 15 14 13 12 11
250000 200000 150000 100000 50000
300000 200000 100000 0
10000 50000
11 13 15
0 300000


balance
60000 50000 40000 30000 20000 10000 0
10 20 30 40 50
age


balance
60000 50000 40000 30000 20000 10000 0
11 12 13 14 15 16 17
education


balance
60000 50000 40000 30000 20000 10000 0
0 100000 200000 300000
homeVal

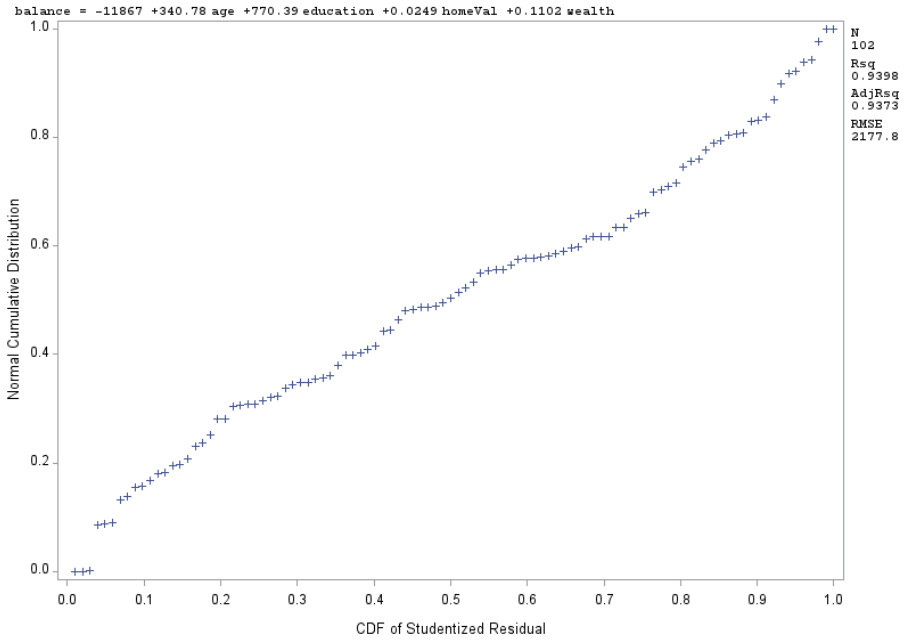

balance
60000 50000 40000 30000 20000 10000 0
0 100000 200000 300000 400000
wealth

2. Plot residuals vs each x-variable.

   None of them show almost a straight line as it was shown in the residuals vs each x-variable graphs in the previous page. Therefore, there is no linearity

Assumptions for Normality:

Plot the normal probability plot of the residuals. It's almost straight which means it's normal.

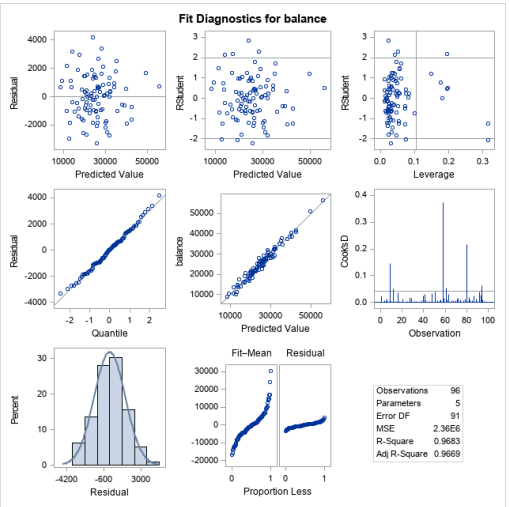


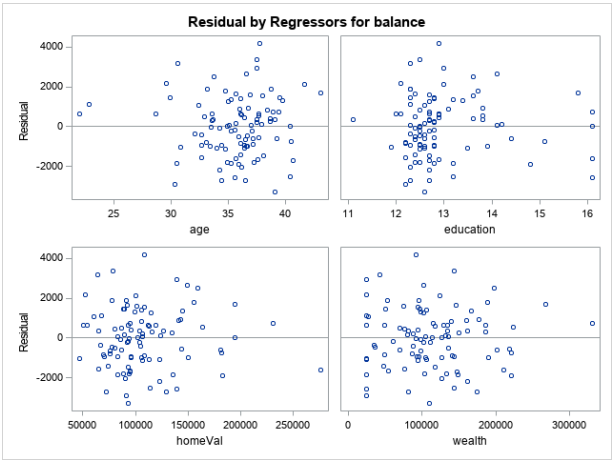
c. Analyze if there are any outliers and/or influential points for your M2 model. If so, what actions would you take to address this issue? Make sure to implement any actions you specify here. Include the relevant output.

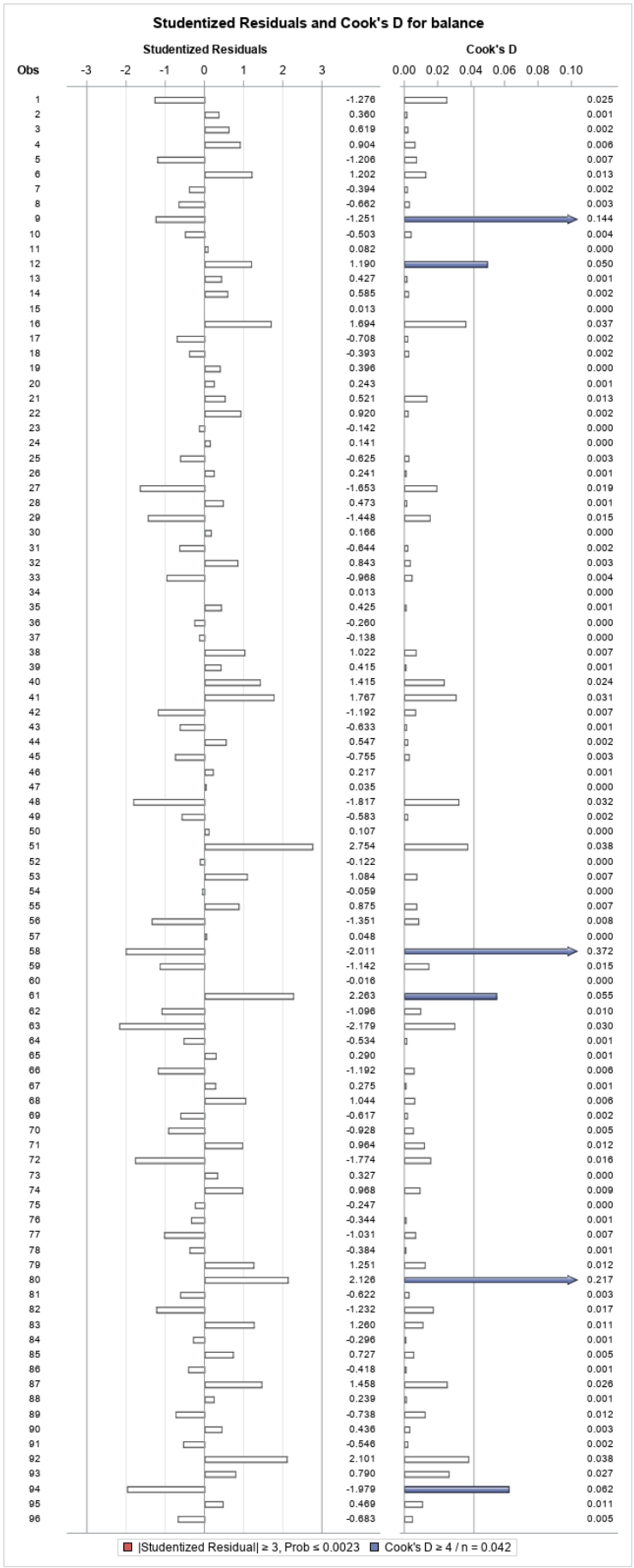
Outliers are the observation numbers (38, 85, 91, 102)

Influential points are the observation numbers (38, 77, 84, 85, 91, 102)

After removing all these 6 observations and rerun the model, there will be no outliers in the new dataset.






Studentized Residuals and Cook's D for balance
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.00 0.02 0.04 0.06 0.08 0.10
1 -1.276 0.025
2 0.360 0.001
3 0.619 0.002
4 0.904 0.006
5 -1.206 0.007
6 1.202 0.013
7 -0.394 0.002
8 -0.662 0.003
9 -1.251 0.144
10 -0.503 0.004
11 0.082 0.000
12 1.190 0.050
13 0.427 0.001
14 0.585 0.002
15 0.013 0.000
16 1.694 0.037
17 -0.708 0.002
18 -0.393 0.002
19 0.396 0.000
20 0.243 0.001
21 0.521 0.013
22 0.920 0.002
23 -0.142 0.000
24 0.141 0.000
25 -0.625 0.003
26 0.241 0.001
27 -1.653 0.019
28 0.473 0.001
29 -1.448 0.015
30 0.166 0.000
31 -0.644 0.002
32 0.843 0.003
33 -0.968 0.004
34 0.013 0.000
35 0.425 0.001
36 -0.260 0.000
37 -0.138 0.000
38 1.022 0.007
39 0.415 0.001
40 1.415 0.024
41 1.767 0.031
42 -1.192 0.007
43 -0.633 0.001
44 0.547 0.002
45 -0.755 0.003
46 0.217 0.001
47 0.035 0.000
48 -1.817 0.032
49 -0.583 0.002
50 0.107 0.000
51 2.754 0.038
52 -0.122 0.000
53 1.084 0.007
54 -0.059 0.000
55 0.875 0.007
56 -1.351 0.008
57 0.048 0.000
58 -2.011 0.372
59 -1.142 0.015
60 -0.016 0.000
61 2.263 0.055
62 -1.096 0.010
63 -2.179 0.030
64 -0.534 0.001
65 0.290 0.001
66 -1.192 0.006
67 0.275 0.001
68 1.044 0.006
69 -0.617 0.002
70 -0.928 0.005
71 0.964 0.012
72 -1.774 0.016
73 0.327 0.000
74 0.968 0.009
75 -0.247 0.000
76 -0.344 0.001
77 -1.031 0.007
78 -0.384 0.001
79 1.251 0.012
80 2.126 0.217
81 -0.622 0.003
82 -1.232 0.017
83 1.260 0.011
84 -0.296 0.001
85 0.727 0.005
86 -0.418 0.001
87 1.458 0.026
88 0.239 0.001
89 -0.738 0.012
90 0.436 0.003
91 -0.546 0.002
92 2.101 0.038
93 0.790 0.027
94 -1.979 0.062
95 0.469 0.011
96 -0.683 0.005
|Studentized Residual| ≥ 3, Prob ≤ 0.0023
Cook's D ≥ 4 / n = 0.042

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

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6571764872	1642941218	695.58	<.0001
Error	91	214940882	2361988		
Corrected Total	95	6786705755			

Root MSE	1536.87596	R-Square	0.9683
Dependent Mean	25444	Adj R-Sq	0.9669
Coeff Var	6.04022		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-14532	3680.07603	-3.95	0.0002	0
age	1	401.30481	54.72409	7.33	<.0001	0.15949
education	1	846.92571	331.58117	2.55	0.0123	0.09588
homeVal	1	0.02396	0.01021	2.35	0.0211	0.10919
wealth	1	0.10696	0.00435	24.58	<.0001	0.74858

Using the absolute value of standardized estimate to determine the predictors with significant effect on balance. The strongest predictor is wealth since the standardized estimate is the highest 0.74858.

Fitted Regression line Expression for M2:
Balance = -14532 + 401.3*age + 846.93*education + 0.024*homeVal + 0.107*wealth + e

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

```
* a;
data Bankingfull;
infile "S:\HW4\Bankingfull.txt" delimiter = '09'x missover firstobs=2;
input age education income homeVal wealth balance;
run;
proc print;
run;

proc sgscatter;
matrix balance age education income homeVal wealth;
run;

*b;
proc corr;
var balance age education income homeVal wealth;
run;

* c;
proc reg;
model balance= age education income homeVal wealth /vif tol;
run;
```

```
data Bankingfull_new;
set Bankingfull;
drop income;
run;

proc print data = Bankingfull_new;
run;

* d;
proc reg;
model balance= age education homeVal wealth /vif tol;
run;

* d-b;
* Residual plot: residuals vs x-variables;
plot student.*(age education homeVal wealth);
* Residual plot: residuals vs pred. values;
plot student.*predicted.;

* Normal probability plot or QQ plot;
plot npp.*student.;
run;

proc sgscatter;
matrix balance age education homeVal wealth ;
run;

proc gplot;
plot balance*(age education homeVal wealth );
run;

proc corr;
var balance age education homeVal wealth;
run;

* d-c;
*run model with outlier - use the second model;
proc reg data = Bankingfull_new;
model balance= age education homeVal wealth/influencec r;
plot student.*( age education homeVal wealth predicted.);
plot npp.*student.;
run;

*deleting Multiple observations;
data bankingfull_new1;
set bankingfull_new;
if _n_ in (38 77 84 85 91 102) then delete;
run;

*rerunning the model without outlier using the new dataset;
proc reg data = Bankingfull_new1;
model balance= age education homeVal wealth/influencec r;
plot student.*( age education homeVal wealth predicted.);
plot npp.*student.;
run;

* d-d;
proc reg data = Bankingfull_new1;
model balance= age education homeVal wealth/stb;
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.

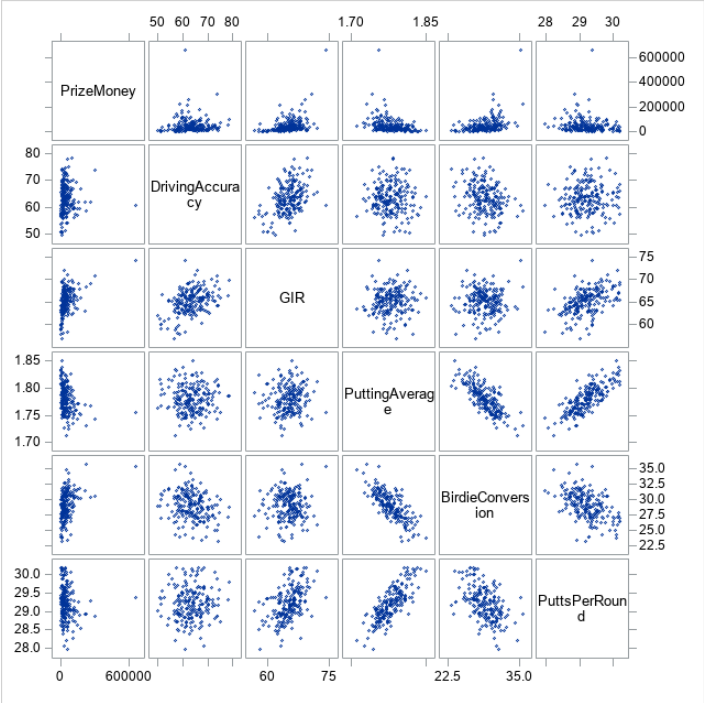
- DrivingAccuracy = percent of times a player is able to hit the fairway with his tee shot
- GIR = percent of time a player was able to hit the green within two or less than par (Greens in Regulation)
- BirdieConversion = percentage of times a player makes a birdie or better after hitting the green in 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)

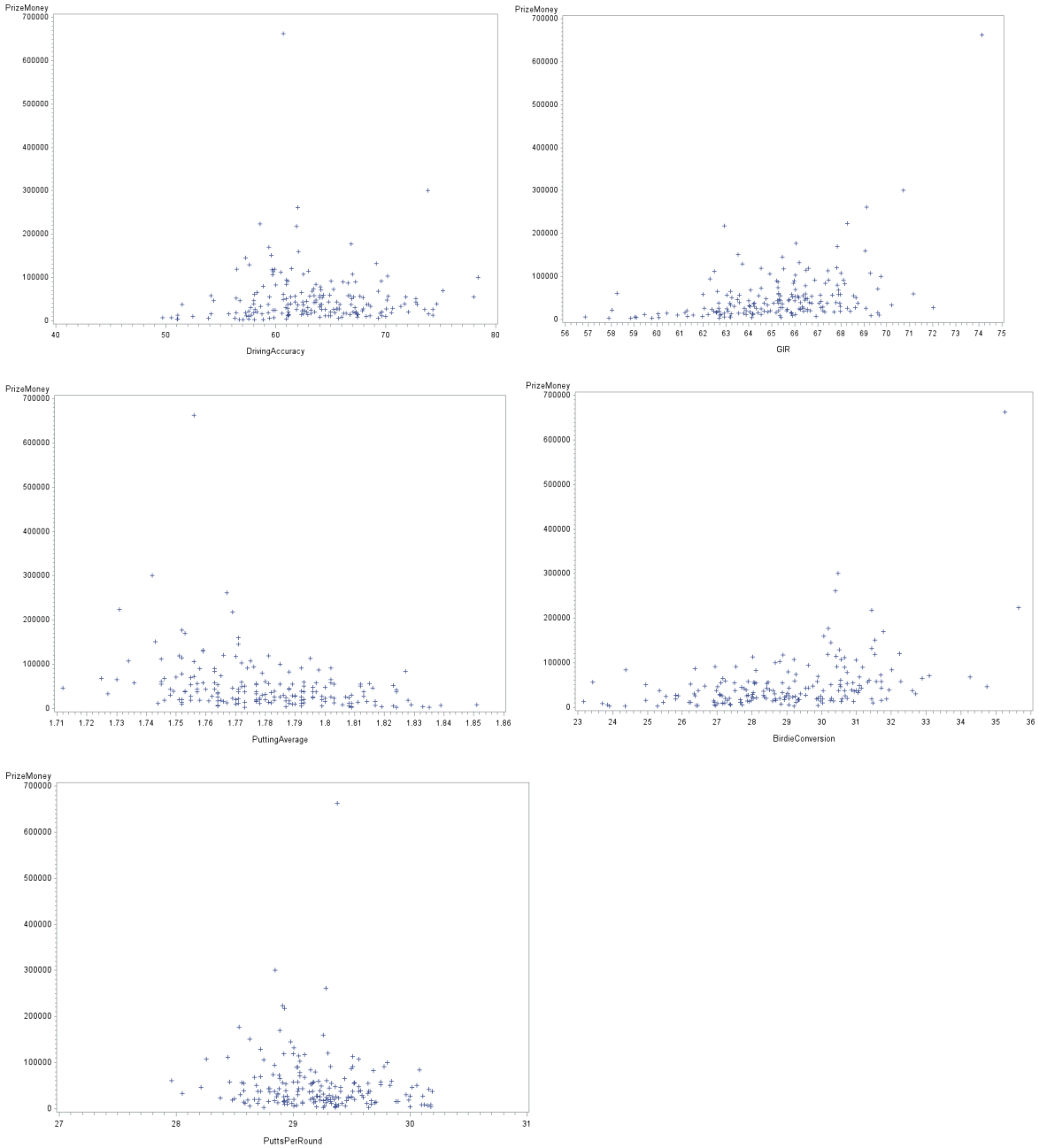
You are asked to build a model for PrizeMoney using the remaining predictors, and to evaluate the relative importance of each different aspects of a player's game on the average prize money.

SAS Code to Import the data

```
*import data from file;
proc import datafile="pgatour2006.csv" out=myd replace;
delimiter=',';
getnames=yes;
run;
```

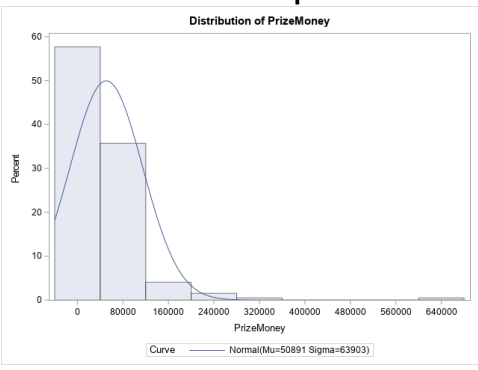
a) Create scatterplots to visualize the associations between PrizeMoney and the other 5 variables. Discuss the patterns displayed by the scatterplot. Also, explain if the associations appear to be linear? (you can create scatterplots or a matrix plot). Include the relevant output.





There is no linear correlation between any of the X-variables and Y-variable.

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



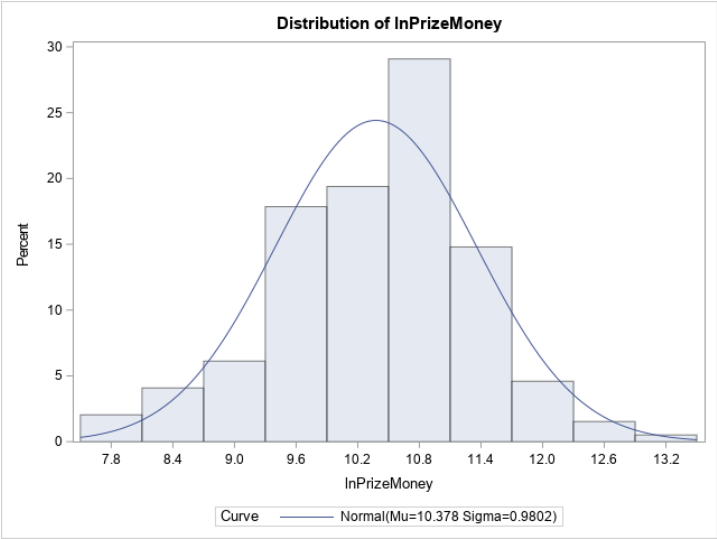
Basic Statistical Measures			
Location		Variability	
Mean	50891.17	Std Deviation	63903
Median	36644.50	Variance	4083587455
Mode	.	Range	660531
		Interquartile Range	40785

Quantiles (Definition 5)	
Level	Quantile
100% Max	662771.0
99%	300555.0
95%	145414.0
90%	107294.0
75% Q3	58006.5
50% Median	36644.5
25% Q1	17221.5
10%	9149.0
5%	5285.0
1%	2426.0
0% Min	2240.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2240	70	217748	63
2426	101	224027	142
2692	97	262045	2
3025	115	300555	90
3635	3	662771	178

The median is $ 36644.5 which means half of the records have a balance that's higher than $36644.5. The first quartile is $ 17221.5 and the third is $ 58006.5 which mean 75% of the records are above $ 17221.5 and 75% of the records are below $ 58006.5.
The distribution is right skewed because the right tail is longer and Mean > Median. It's positively skewed and most of the values fall within the lower range.

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.



Quantiles (Definition 5)	
Level	Quantile
100% Max	13.40418
99%	12.61339
95%	11.88734
90%	11.58333
75% Q3	10.96831
50% Median	10.50900
25% Q1	9.75377
10%	9.12140
5%	8.57263
1%	7.79400
0% Min	7.71423

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
7.71423	70	12.2911	63
7.79400	101	12.3195	142
7.89804	97	12.4763	2
8.01467	115	12.6134	90
8.19836	3	13.4042	178

Basic Statistical Measures			
Location		Variability	
Mean	10.37808	Std Deviation	0.98020
Median	10.50900	Variance	0.96080
Mode	.	Range	5.68995
		Interquartile Range	1.21454

Moments			
N	196	Sum Weights	196
Mean	10.3780793	Sum Observations	2034.10355
Std Deviation	0.98020183	Variance	0.96079563
Skewness	-0.2017064	Kurtosis	0.24484282
Uncorrected SS	21297.4432	Corrected SS	187.355147
Coeff Variation	9.44492519	Std Error Mean	0.07001442

After we applied log transformation, the distribution became almost normal since the mean is approximately equal to the median. Also, the histogram curve is symmetrical.

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.

a) 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.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	8.24102	7.16112	1.15	0.2513	0
DrivingAccuracy	1	-0.00075836	0.01161	-0.07	0.9480	-0.00419
GIR	1	0.26879	0.02879	9.33	<.0001	0.74652
PuttingAverage	1	8.74678	5.37342	1.63	0.1052	0.22066
BirdieConversion	1	0.15230	0.04083	3.73	0.0003	0.34285
PuttsPerRound	1	-1.20948	0.26728	-4.53	<.0001	-0.54502

The strongest predictor is GIR since the absolute value of standardized estimate is the highest 0.7465. When performing t-test on individual parameters, GIR, BirdieConversion, and PuttsPerRound have p-values that are less than 0.05 which make them significant X variables. The DrivingAccuracy has the highest p-value 0.948 which makes it insignificant.

Next step is to remove DrivingAccuracy variable and rerun the model again.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	8.02738	6.35383	1.26	0.2080	0
GIR	1	0.26791	0.02536	10.56	<.0001	0.74407
PuttingAverage	1	8.81065	5.26991	1.67	0.0962	0.22227
BirdieConversion	1	0.15360	0.03561	4.31	<.0001	0.34576
PuttsPerRound	1	-1.20702	0.26391	-4.57	<.0001	-0.54391

After we rerun the model, the PuttingAverage has the highest p-value 0.096 that's higher than 0.05 which makes it insignificant.
Next step is to remove PuttingAverage variable and rerun the model again.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	15.81016	4.34465	3.64	0.0004	0
GIR	1	0.24542	0.02160	11.36	<.0001	0.68162
BirdieConversion	1	0.11454	0.02700	4.24	<.0001	0.25785
PuttsPerRound	1	-0.84757	0.15377	-5.51	<.0001	-0.38193

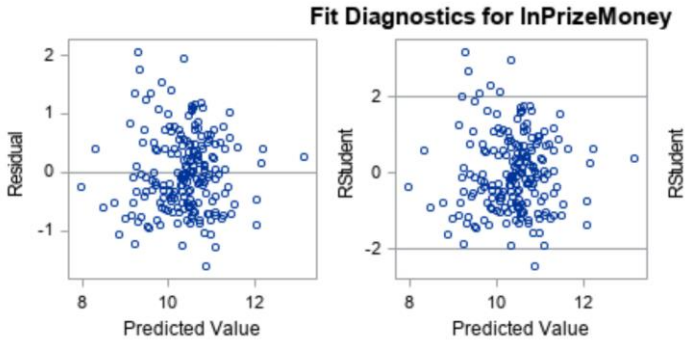
The Fitted Regression Model:

lnPrizeMoney = 15.81 + 0.24542*GIR + 0.11454*BirdieConversion -0.84757*PuttsPerRound + e

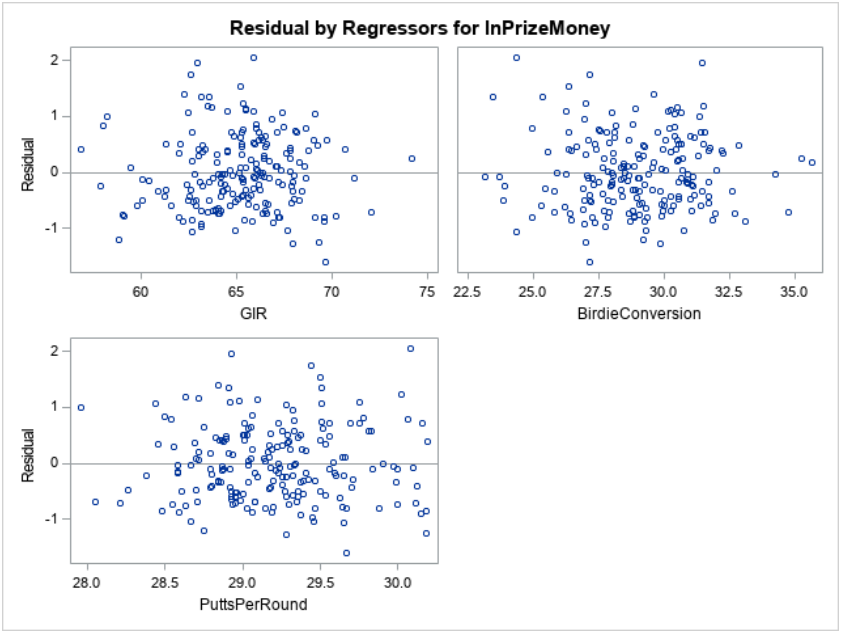
b) Analyze residual plots to check if the regression model is valid for your data. Discuss your analysis.

Assumptions of Constant Variance and Independence:

1. Plot residuals vs predicted values.



2. Plot residuals vs each x-variable.

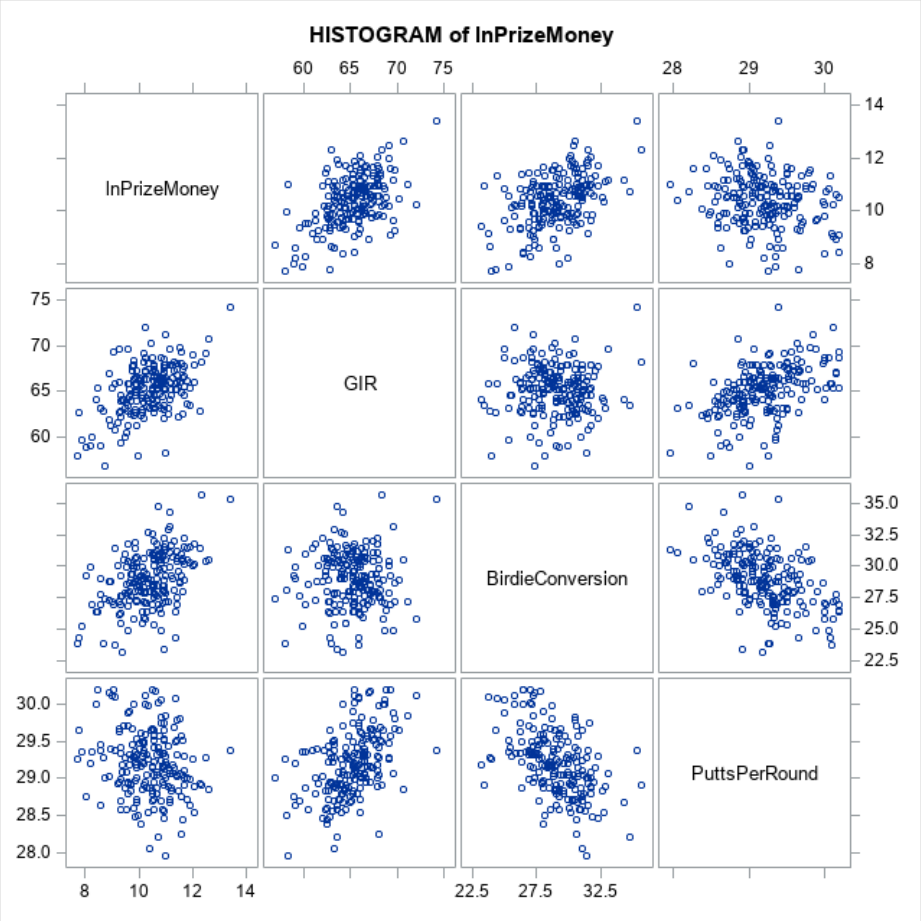


Points are randomly scattered, and residual analysis show no concern for the model fit.

Assumptions for linearity:

1. Scatterplot for each x-variable.

   There is no linearity shown in all scatterplots where Y-variable is plotted against each X-variable as it's shown in the next matrix scatterplot.

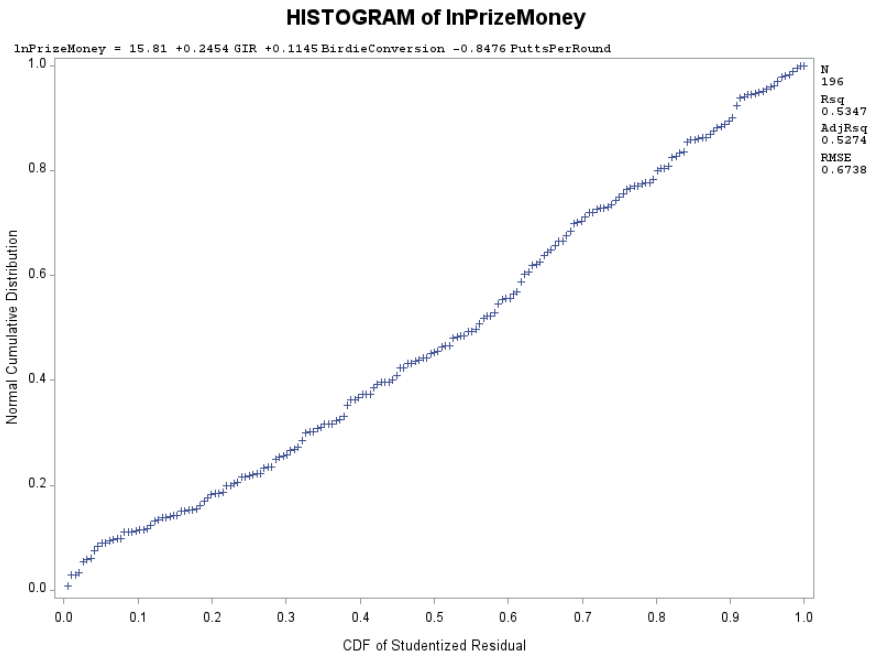


2. Plot residuals vs each x-variable.

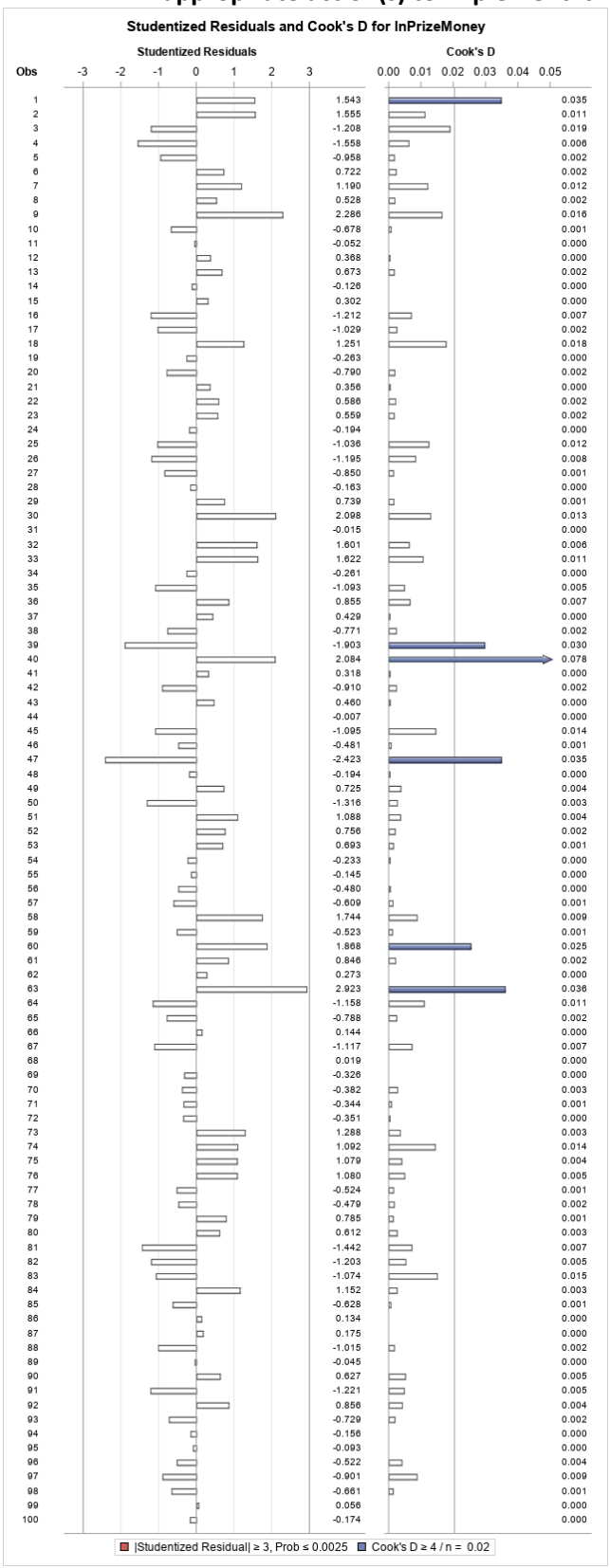
   There is no linearity shown in the previous graph where residuals are plotted against each x-variable.

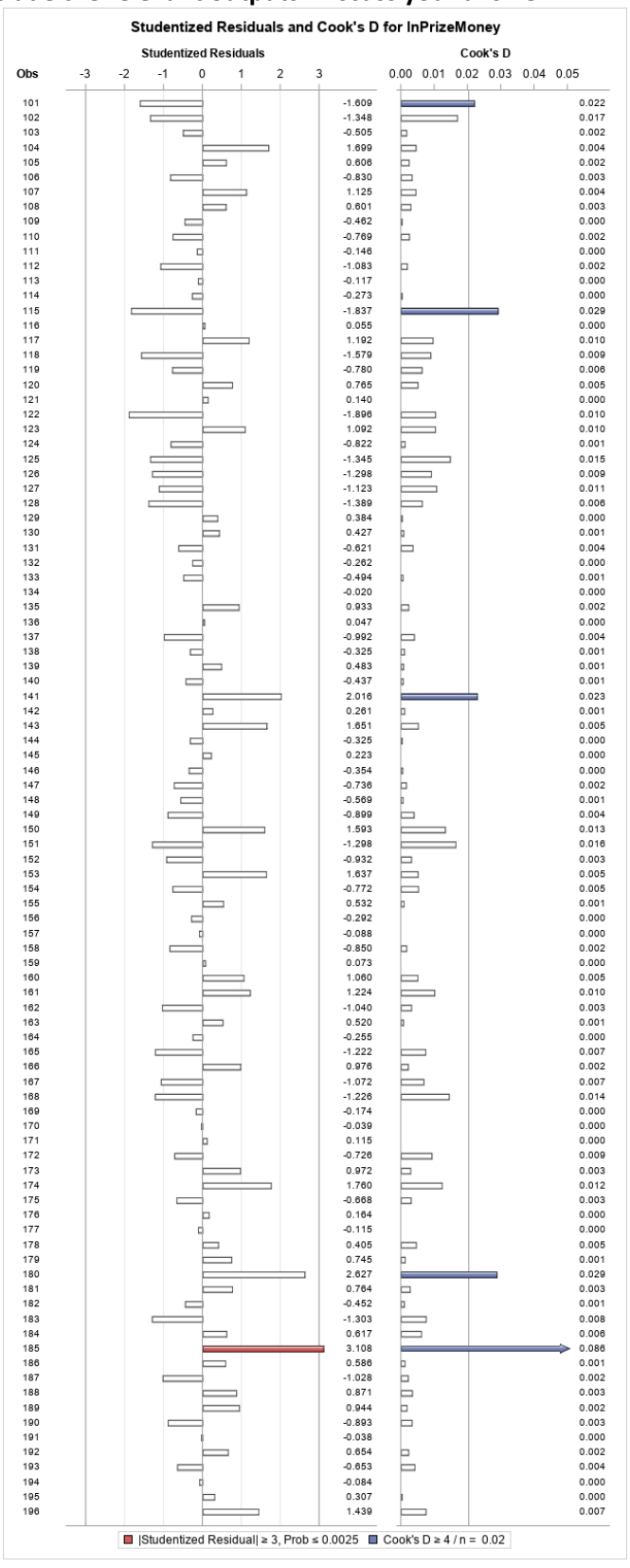
Assumptions for Normality:

Plot the normal probability plot of the residuals. It's almost a straight line which makes it normal.



i. 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.





After removing observations number (40, 185), we rerun the model again.

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

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.73166	4.29460	3.43	0.0007
GIR	1	0.24409	0.02095	11.65	<.0001
BirdieConversion	1	0.13566	0.02701	5.02	<.0001
PuttsPerRound	1	-0.82926	0.15137	-5.48	<.0001

lnPrizeMoney = 14.73 + 0.24409*GIR + 0.13566*BirdieConversion -0.82926*PuttsPerRound + e

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?

e(0.24)-1 = 0.27 = 27%

lnPrizeMoney increases by 27 % when GIR increases 1%

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

```
*import data from file;
proc import datafile="S:\HW4\pgatour2006.csv" out=myd replace;
delimiter=',';
getnames=yes;
run;

Proc print;
run;

* a;
proc sgscatter;
matrix PrizeMoney DrivingAccuracy GIR PuttingAverage BirdieConversion
PuttsPerRound;
run;

proc gplot;
plot PrizeMoney*(DrivingAccuracy GIR PuttingAverage BirdieConversion
PuttsPerRound);
run;

proc corr;
var PrizeMoney DrivingAccuracy GIR PuttingAverage BirdieConversion
PuttsPerRound;
run;

* b;
title "HISTOGRAM of PrizeMoney";
proc univariate normal;
var PrizeMoney;
histogram / normal (mu = est sigma = est);
run;

data myd;
```

```
set myd;
lnPrizeMoney = log(PrizeMoney);
run;
Proc print;
run;

* c;
title "HISTOGRAM of lnPrizeMoney";
proc univariate normal;
var lnPrizeMoney;
histogram / normal (mu = est sigma = est);
run;

proc reg;
model lnPrizeMoney= DrivingAccuracy GIR PuttingAverage BirdieConversion
PuttsPerRound /stb;
run;

* d-a;
*rerunning the model after removing insig DrivingAccuracy;
 proc reg;
model lnPrizeMoney= GIR PuttingAverage BirdieConversion PuttsPerRound /stb;
run;

*rerunning the model again after removing insig DrivingAccuracy &
PuttingAverage;
 proc reg;
model lnPrizeMoney= GIR BirdieConversion PuttsPerRound /stb;
run;

* d-b;
* Residual plot: residuals vs x-variables;
plot student.*(GIR BirdieConversion PuttsPerRound);
* Residual plot: residuals vs pred. values;
plot student.*predicted.;

* Normal probability plot or QQ plot;
plot npp.*student.;
run;

proc sgscatter;
matrix lnPrizeMoney GIR BirdieConversion PuttsPerRound ;
run;

proc gplot;
plot lnPrizeMoney*(GIR BirdieConversion PuttsPerRound );
run;

proc corr;
var lnPrizeMoney GIR BirdieConversion PuttsPerRound;
run;

*run model with outlier ;
proc reg data = myd;
model lnPrizeMoney=GIR BirdieConversion PuttsPerRound/influencec r;
plot student.*( GIR BirdieConversion PuttsPerRound predicted.);
plot npp.*student.;
run;

*deleting Multiple observations;
```

```
data myd_new;
set myd;
if _n_ in (40 185) then delete;
run;

*rerun model use the second model;
proc reg data = myd_new;
model lnPrizeMoney=GIR BirdieConversion PuttsPerRound/influencec r;
plot student.*( GIR BirdieConversion PuttsPerRound predicted.);
plot npp.*student.;
run;

proc reg data = myd_new;
model lnPrizeMoney=GIR BirdieConversion PuttsPerRound;
run;
```