XIMAN LIU

DATA ANALYSIS AND REGRESSION

Assignment-4 | Total Points: 46 pts for DSC 423 and 26 pts for DSC 323
Due Date: 4/26/2021 by 11:59 pm

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 (8-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.

***Note:* For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.**

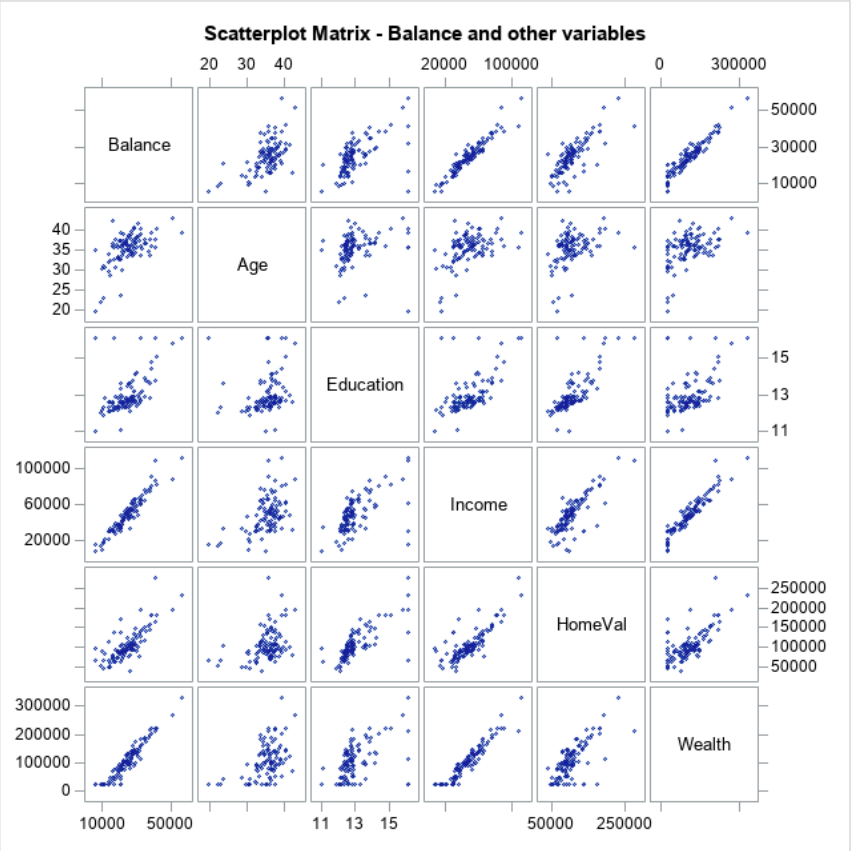
PROBLEM 1 [16 pts] – to be answered by everyone

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

- median age of the population (AGE)
- median years of education (EDUCATION)
- median income (INCOME) in $
- median home value (HOMEVAL) in $
- median household wealth (WEALTH) in $
- average bank balance (BALANCE) in $

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

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)



The graph represents positive linear relationship between balance and other variables (age, education, income, homeval and wealth). The relationship between balance and income, and the relationship between balance and wealth represent strong associations. There are possible potential outliers at the top right corner.
The relationship of age and balance is weak association because there is a cloud in the graph.
The relationship of education and balance is weak association because there is a cloud in the graph.
The relationship of homeval and balance is strong association. There are possible potential outliers in the plot.

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

The higher the correlation r value (between 0 and 1), the stronger the linear the linear relationship.
r(balance, income) = 0.95168 and r(balance, wealth) = 0.94871 represent highly strong perfect positive association between balance and income, and balance and wealth.
r(balance, homeval) = 0.76639 represents a relatively strong association between balance and homeval.
r(balance, age) = 0.56547 and r(balance, education) = 0.55488 represent a moderate weak association between balance and age, and balance 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.

The REG Procedure
Model: MODEL1
Dependent Variable: Balance

Number of Observations Read	102
Number of Observations Used	102

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|	Variance Inflation
Intercept	1	-10711	4260.97631	-2.51	0.0136	0
Age	1	318.66496	60.98611	5.23	<.0001	1.34276
Education	1	621.86035	318.95952	1.95	0.0541	2.45671
Income	1	0.14632	0.04078	3.59	0.0005	14.90172
HomeVal	1	0.00918	0.01104	0.83	0.4075	4.38300
Wealth	1	0.07433	0.01119	6.64	<.0001	10.71428

Multicollinearity is when two or more of the predictors in a regression model are moderately or highly correlated with each other. If VIF of a variable is greater than 10, then it is considered to be multicollinearity. The R-Square is 0.9469 which is a quite high number. In the graph, income and wealth are both greater than 10, which means high possibility of multicollinearity. Hence, the estimation of regression coefficients is weak.
The corresponding solution is, remove either income or wealth from the model. The reason for removing income is, income might not be a sufficient predictor for balance. Vice versa, we could also remove wealth here, because holders with a lot of money might have properties that aren't recorded in the bank.

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:

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

Drop Wealth

Regression Model - Balance and Other Variables New 1

The REG Procedure
Model: MODEL1
Dependent Variable: Balance

Number of Observations Read	102
Number of Observations Used	102

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7048697167	1762174292	288.66	<.0001
Error	97	592146978	6104608		
Corrected Total	101	7640844145			

Root MSE	2470.75050	R-Square	0.9225
Dependent Mean	24888	Adj R-Sq	0.9193
Coeff Var	9.92752		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-9809.37834	5118.82163	-1.92	0.0583	0
Age	1	333.45082	73.25253	4.55	<.0001	1.34098
Education	1	313.00839	379.27426	0.83	0.4112	2.40451
Income	1	0.38845	0.02199	17.67	<.0001	2.99818
HomeVal	1	-0.00139	0.01313	-0.11	0.9157	4.29181

Drop Income (M2)

Residual Plots - Balance and other variables New 2)

The REG Procedure
Model: MODEL1
Dependent Variable: Balance

Number of Observations Read	102
Number of Observations Used	102

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			

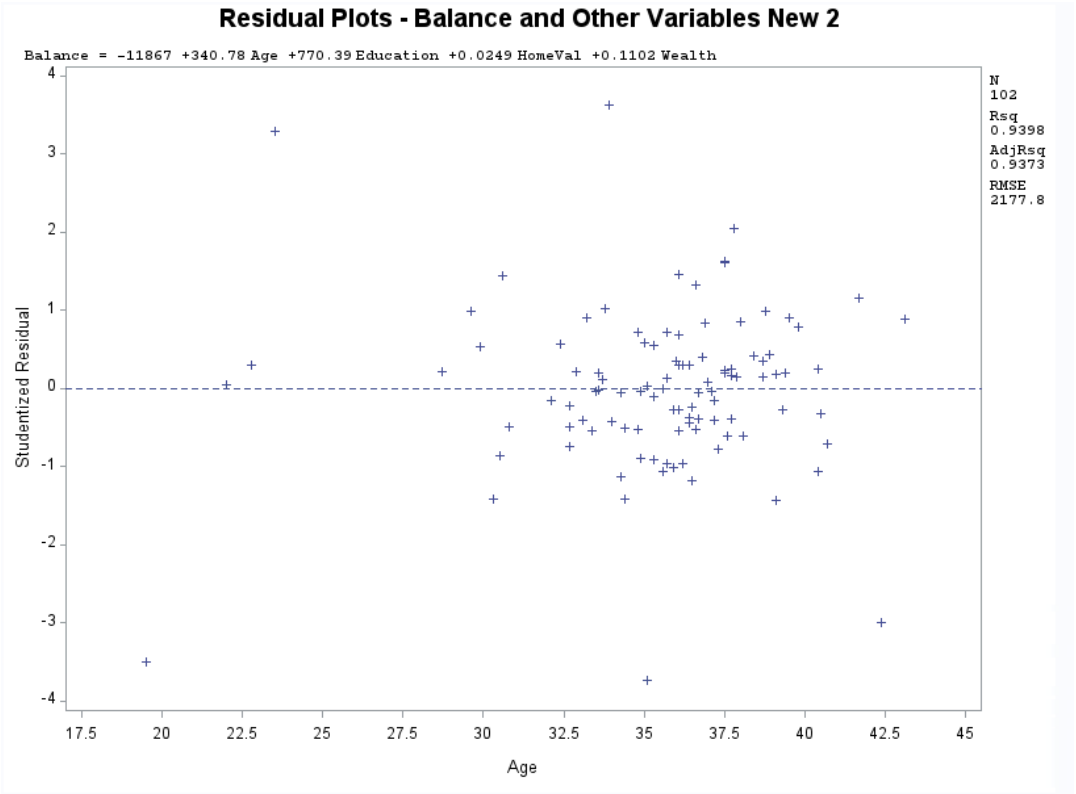
Root MSE	2177.82964	R-Square	0.9398
Dependent Mean	24888	Adj R-Sq	0.9373
Coeff Var	8.75056		

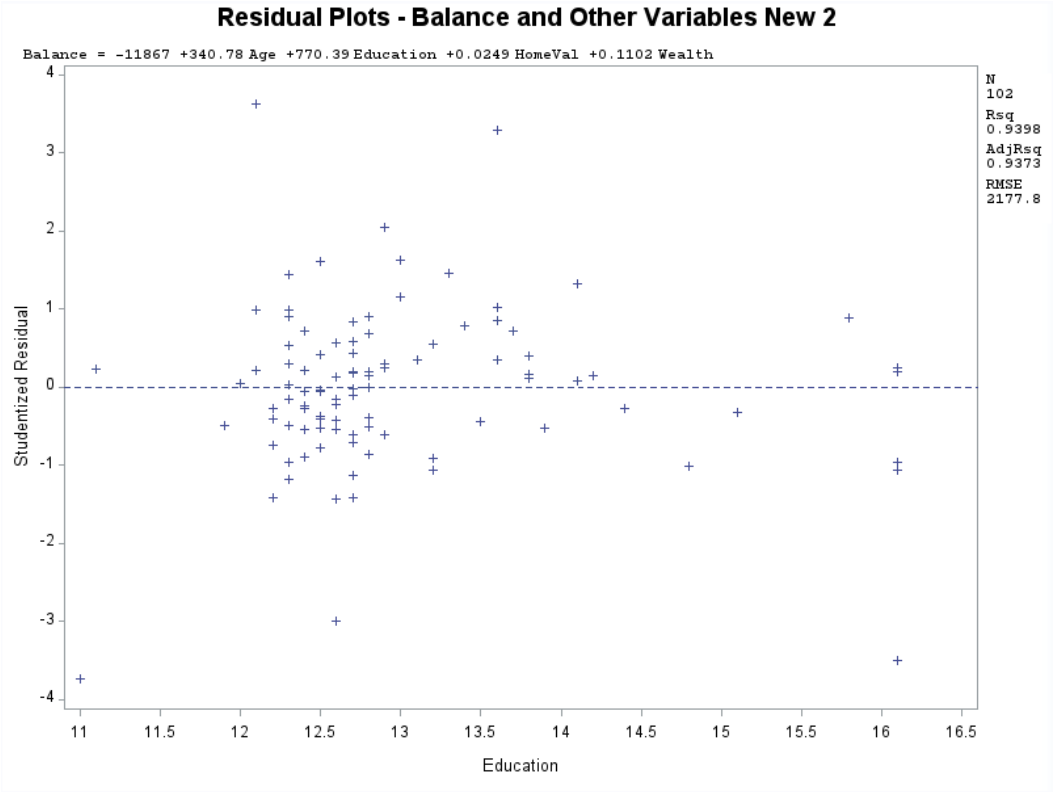
Parameter Estimates

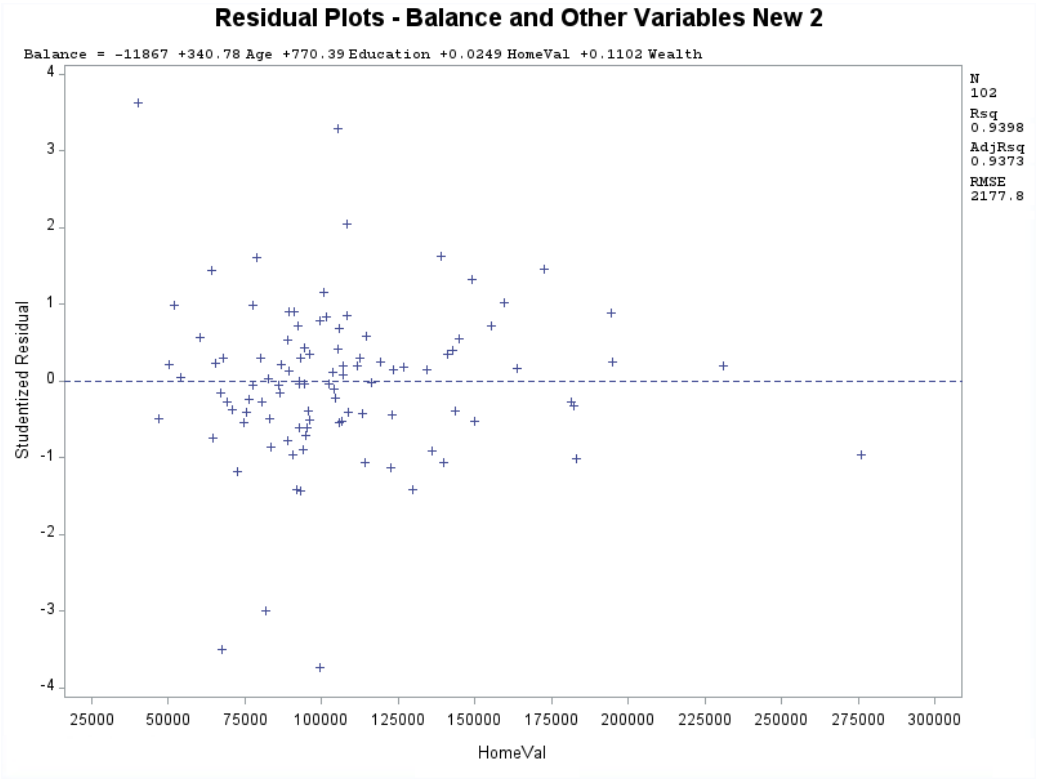
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-11867	4501.31282	-2.64	0.0098	0
Age	1	340.77546	64.28041	5.30	<.0001	1.32905
Education	1	770.39302	335.06016	2.30	0.0236	2.41532
HomeVal	1	0.02485	0.01074	2.31	0.0228	3.69684
Wealth	1	0.11021	0.00532	20.73	<.0001	2.15568

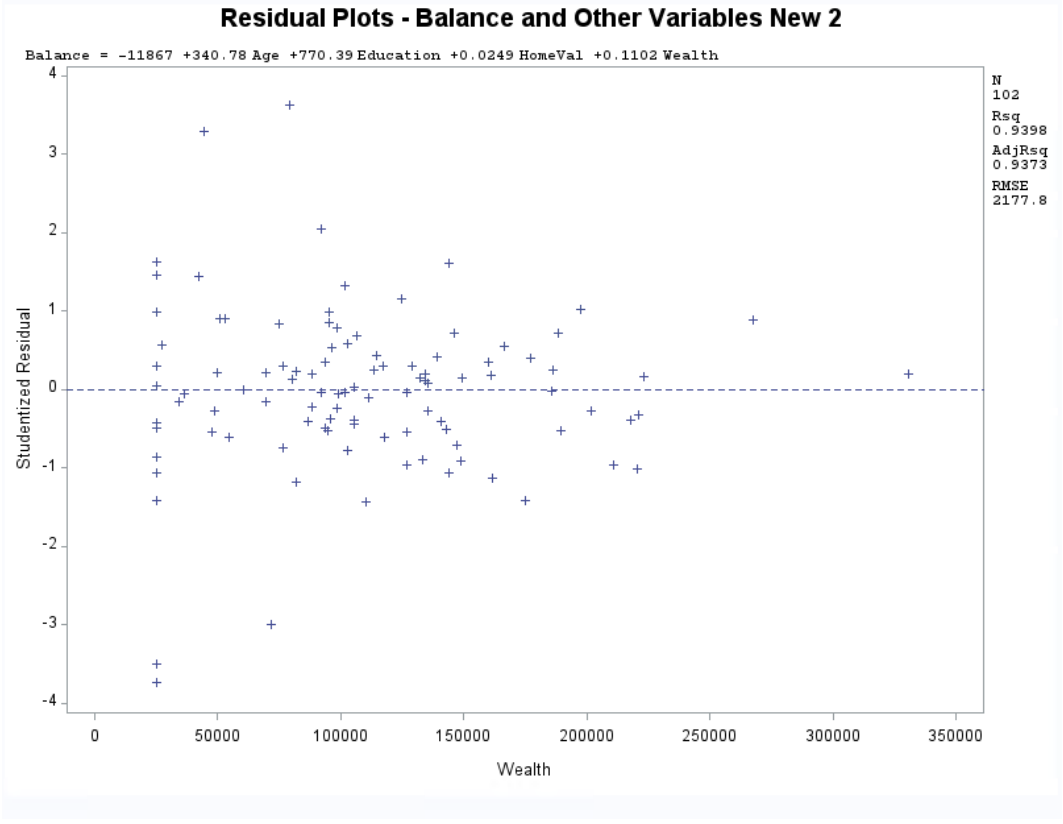
Compared the dataset removing income and the one removing wealth, we can tell that the regression model removing income has a higher R-Square (0.9398) and adj R-Square (0.9373). In that case, we choose the model removing income (M2).

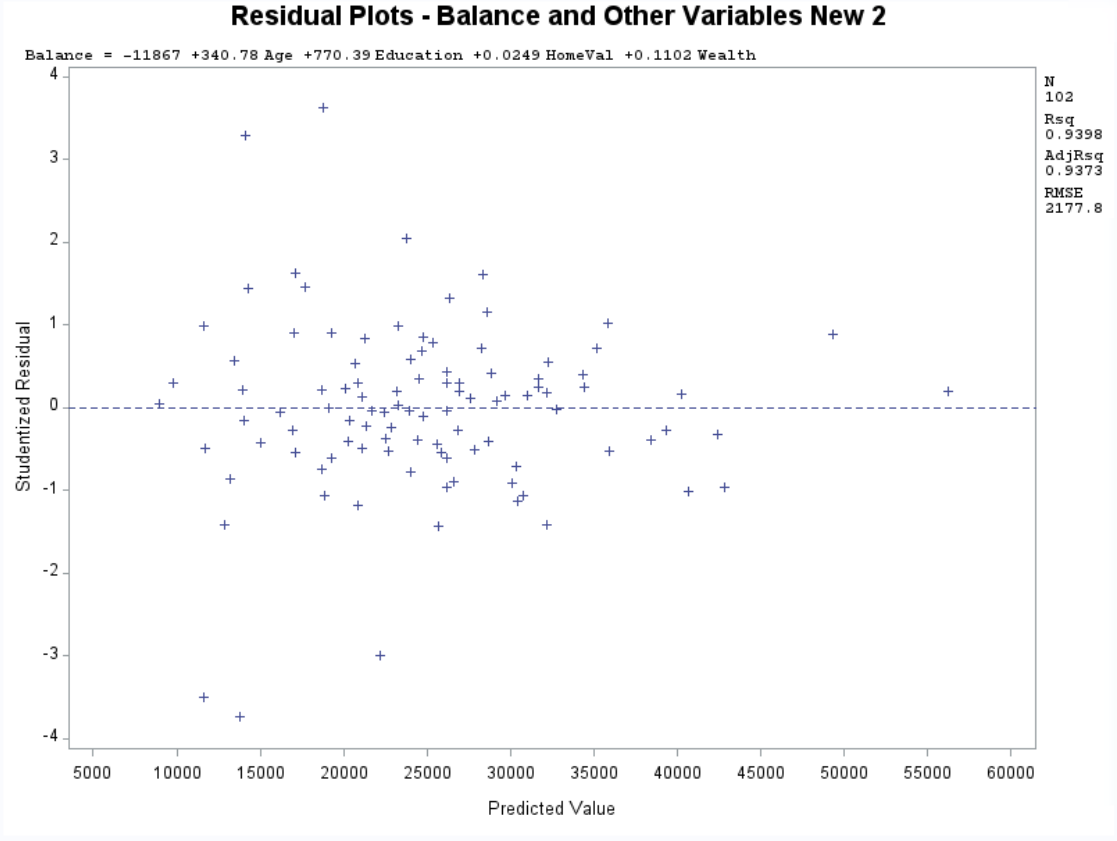
b. Create residual plots for M2 (Studentized residuals vs predicted; Studentized residuals vs x-variables; and normal plot of residuals). Analyze the residual plots to check if the regression model assumptions are met by the data. Include the relevant output and discuss your analysis.

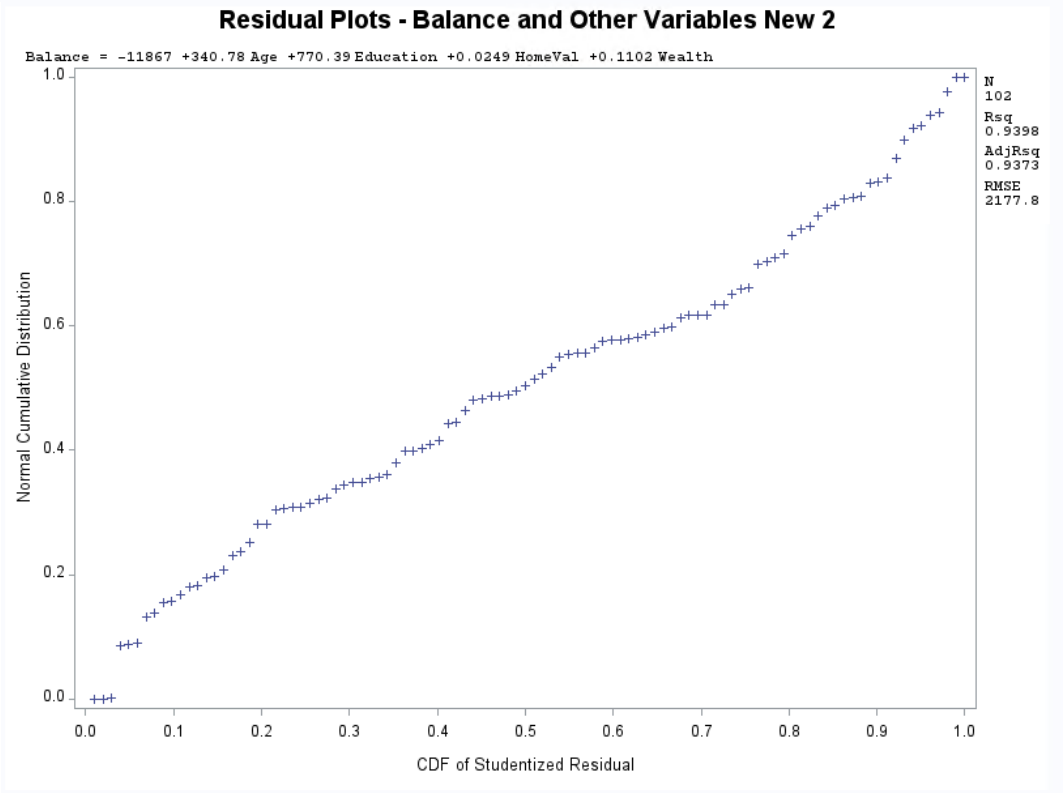

Residual Plots - Balance and Other Variables New 2
Balance = -11867 +340.78 Age +770.39 Education +0.0249 HomeVal +0.1102 Wealth
N
102
Rsq
0.9398
AdjRsq
0.9373
RMSE
2177.8
Studentized Residual
Age


Residual Plots - Balance and Other Variables New 2
Balance = -11867 +340.78 Age +770.39 Education +0.0249 HomeVal +0.1102 Wealth
N
102
Rsq
0.9398
AdjRsq
0.9373
RMSE
2177.8
Studentized Residual
Education


Residual Plots - Balance and Other Variables New 2
Balance = -11867 +340.78 Age +770.39 Education +0.0249 HomeVal +0.1102 Wealth
N
102
Rsq
0.9398
AdjRsq
0.9373
RMSE
2177.8
Studentized Residual
HomeVal
4
3
2
1
0
-1
-2
-3
-4
25000
50000
75000
100000
125000
150000
175000
200000
225000
250000
275000
300000


Residual Plots - Balance and Other Variables New 2
Balance = -11867 +340.78 Age +770.39 Education +0.0249 HomeVal +0.1102 Wealth
N
102
Rsq
0.9398
AdjRsq
0.9373
RMSE
2177.8
Studentized Residual
Wealth
4
3
2
1
0
-1
-2
-3
-4
0
50000
100000
150000
200000
250000
300000
350000





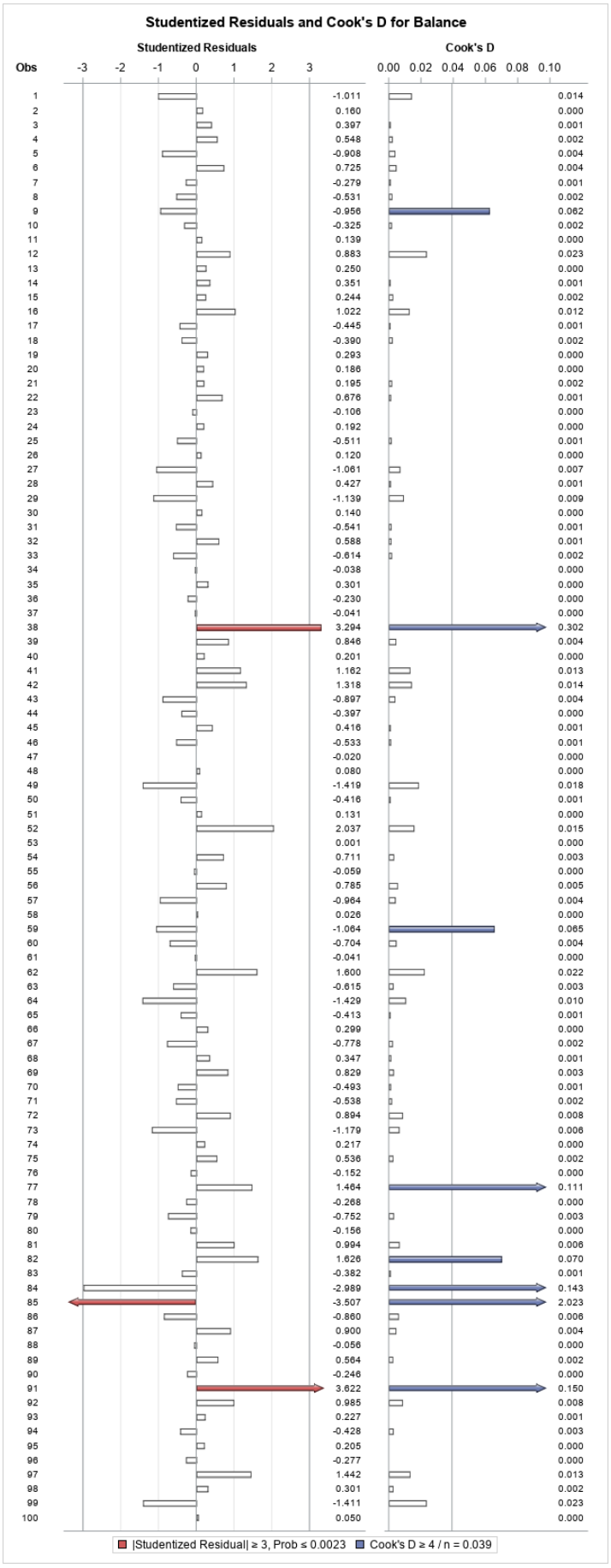
The first five graphs (age, education, home value, wealth and predicted value) have scattered around zero line randomly. And it represents independence and constant variance.

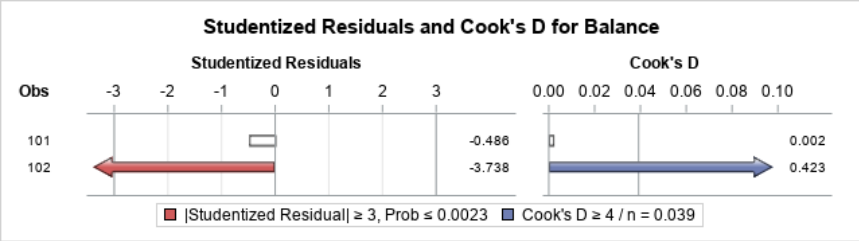
The first four graphs (age, education, home value, wealth) are linearly.

The last one shows a linearly line and it is normal distributed.

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.

Root MSE	1518.96618	R-Square	0.9688
Dependent Mean	25483	Adj R-Sq	0.9675
Coeff Var	5.96076		


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 -1.011 0.014
2 0.160 0.000
3 0.397 0.001
4 0.548 0.002
5 -0.908 0.004
6 0.725 0.004
7 -0.279 0.001
8 -0.531 0.002
9 -0.956 0.062
10 -0.325 0.002
11 0.139 0.000
12 0.883 0.023
13 0.250 0.000
14 0.351 0.001
15 0.244 0.002
16 1.022 0.012
17 -0.445 0.001
18 -0.390 0.002
19 0.293 0.000
20 0.186 0.000
21 0.195 0.002
22 0.676 0.001
23 -0.106 0.000
24 0.192 0.000
25 -0.511 0.001
26 0.120 0.000
27 -1.061 0.007
28 0.427 0.001
29 -1.139 0.009
30 0.140 0.000
31 -0.541 0.001
32 0.588 0.001
33 -0.614 0.002
34 -0.038 0.000
35 0.301 0.000
36 -0.230 0.000
37 -0.041 0.000
38 3.294 0.302
39 0.846 0.004
40 0.201 0.000
41 1.162 0.013
42 1.318 0.014
43 -0.897 0.004
44 -0.397 0.000
45 0.416 0.001
46 -0.533 0.001
47 -0.020 0.000
48 0.080 0.000
49 -1.419 0.018
50 -0.416 0.001
51 0.131 0.000
52 2.037 0.015
53 0.001 0.000
54 0.711 0.003
55 -0.059 0.000
56 0.785 0.005
57 -0.964 0.004
58 0.026 0.000
59 -1.064 0.065
60 -0.704 0.004
61 -0.041 0.000
62 1.600 0.022
63 -0.615 0.003
64 -1.429 0.010
65 -0.413 0.001
66 0.299 0.000
67 -0.778 0.002
68 0.347 0.001
69 0.829 0.003
70 -0.493 0.001
71 -0.538 0.002
72 0.894 0.008
73 -1.179 0.006
74 0.217 0.000
75 0.536 0.002
76 -0.152 0.000
77 1.464 0.111
78 -0.268 0.000
79 -0.752 0.003
80 -0.156 0.000
81 0.994 0.006
82 1.626 0.070
83 -0.382 0.001
84 -2.989 0.143
85 -3.507 2.023
86 -0.860 0.006
87 0.900 0.004
88 -0.056 0.000
89 0.564 0.002
90 -0.246 0.000
91 3.622 0.150
92 0.985 0.008
93 0.227 0.001
94 -0.428 0.003
95 0.205 0.000
96 -0.277 0.000
97 1.442 0.013
98 0.301 0.002
99 -1.411 0.023
100 0.050 0.000
|Studentized Residual| ≥ 3, Prob ≤ 0.0023
Cook's D ≥ 4 / n = 0.039



There are four outliers at obs 38, 85, 91, 102 marked as red color, they are over +3 or -3 range. Besides, the influential points marked as blue are obs 9, 38, 59, 77, 82, 84, 85, 91 and 102.
Firstly, let's say obs 102 is marked with both outlier and influential point. We will remove it. Then we retry it and see whether it helps to improve the result until nothing will be changed at the end.
All in all, I tried removing 6 obs and see the changes on R-Square and adj R-Square.
At this time, with R-Square is 0.9688 and adj R-Square is 0.9675, we can tell that we can stop deleting and this is the final model.

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

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-17098	3983.66356	-4.29	<.0001	0
Age	1	403.81740	53.90071	7.49	<.0001	0.16109
Education	1	1062.99272	357.67921	2.97	0.0038	0.11382
HomeVal	1	0.02636	0.00919	2.87	0.0052	0.12192
Wealth	1	0.10230	0.00399	25.66	<.0001	0.71863

The largest absolute value of the standardized coefficient represents the x-variable that has the greatest influence on Y, and can be considered the strongest predictor of Y.
The strongest influence on balance is wealth, it has highest standardized estimate as 0.71863.

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

```
TITLE "Analysis of Bankingfull";

PROC IMPORT datafile="C:\Users\XLIU115\Desktop\Assignment4\Bankingfull.txt"
out=Balance replace;
getnames=yes;
delimiter='09'x;
RUN;

PROC PRINT data=Balance;
RUN;

/*1A*/
PROC SGSCATTER;
```

```
  title "Scatterplot Matrix - Balance and Other Variables";
  matrix Balance Age Education Income HomeVal Wealth;
RUN;

/*1B*/
PROC CORR;
var Balance Age Education Income HomeVal Wealth;
RUN;


/*1C*/
PROC REG;
title "Regression Model - Balance and Other Variables";
model Balance= Age Education Income HomeVal Wealth /vif;
RUN;


/*1D*/
/*Drop Wealth*/
DATA Balance_1;
set Balance;
drop Wealth;
RUN;

PROC PRINT data= Balance_1;
RUN;

PROC REG;
title "Regression Model - Balance and Other Variables New 1";
model Balance= Age Education Income HomeVal /vif;
RUN;

/*Drop Income*/
DATA Balance_2;
set Balance;
drop Income;
RUN;

PROC PRINT data= Balance_2;
RUN;

PROC REG;
title "Residual Plots - Balance and Other Variables New 2";
model Balance= Age Education HomeVal Wealth /vif;
plot student.*(Age Education HomeVal Wealth predicted.);
plot npp.*student.;
RUN;


PROC REG;
title "Outliers - Balance and Other Variables New 2";
model Balance= Age Education HomeVal Wealth /influence r;
plot student.*(Age Education HomeVal Wealth predicted.);
plot npp.*student.;
RUN;
```

```
DATA Balance_3;
set Balance_3;
if _n_=58 then delete;
RUN;

PROC REG;
title "Regression Model - Balance and other variables New 2";
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.

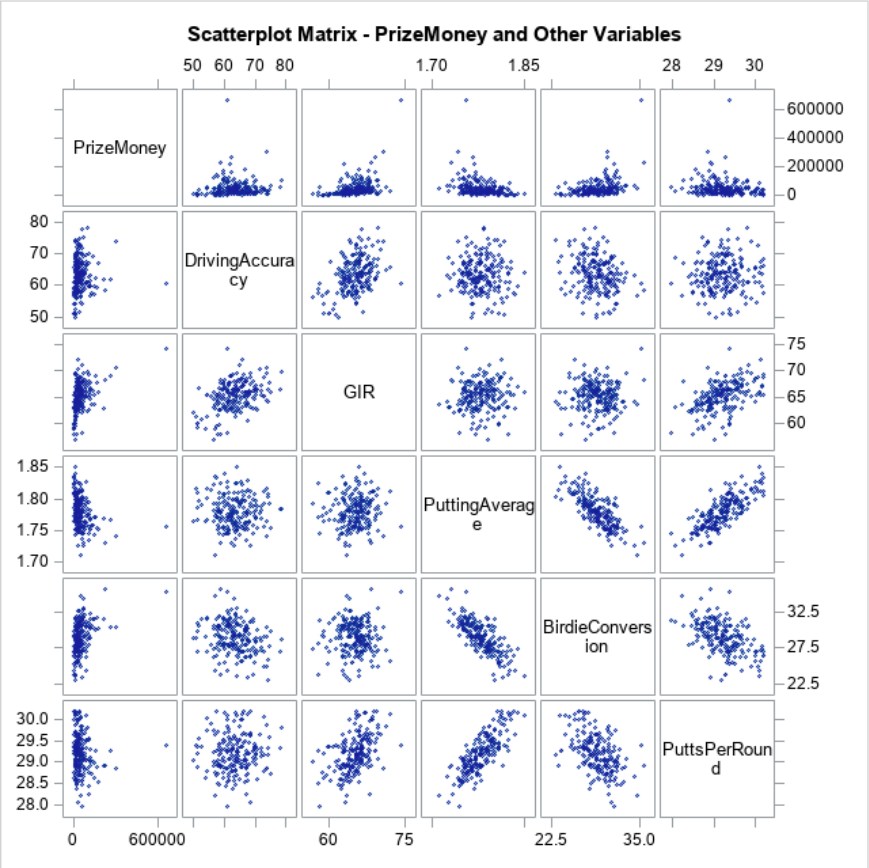
Note: For the non-golfers in the class, you can refer to this page for an explanation of the terms:
http://en.wikipedia.org/wiki/Glossary_of_golf

SAS Code to Import the data

```
*import data from file;
proc import datafile="pgatour2006.csv" out=PGATour 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 `PGATour`. You can change the name if you like.

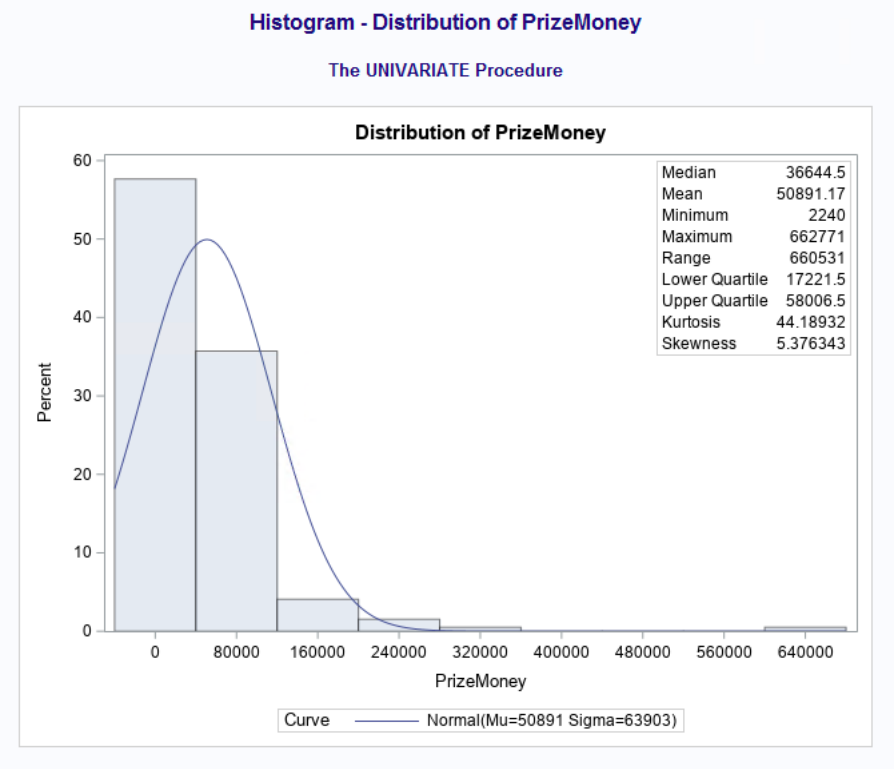
a) Create scatterplots to visualize the associations between PrizeMoney and the other 5 variables. Discuss the patterns displayed by the scatterplot. Also, explain if the associations appear to be linear? (you can create scatterplots or a matrix plot). Include the relevant output.



Pearson Correlation Coefficients, N = 196 Prob > |r| under H0: Rho=0						
	PrizeMoney	DrivingAccuracy	GIR	PuttingAverage	BirdieConversion	PuttsPerRound
PrizeMoney	1.00000	0.02468 0.7314	0.41022 <.0001	-0.31305 <.0001	0.41343 <.0001	-0.11249 0.1165

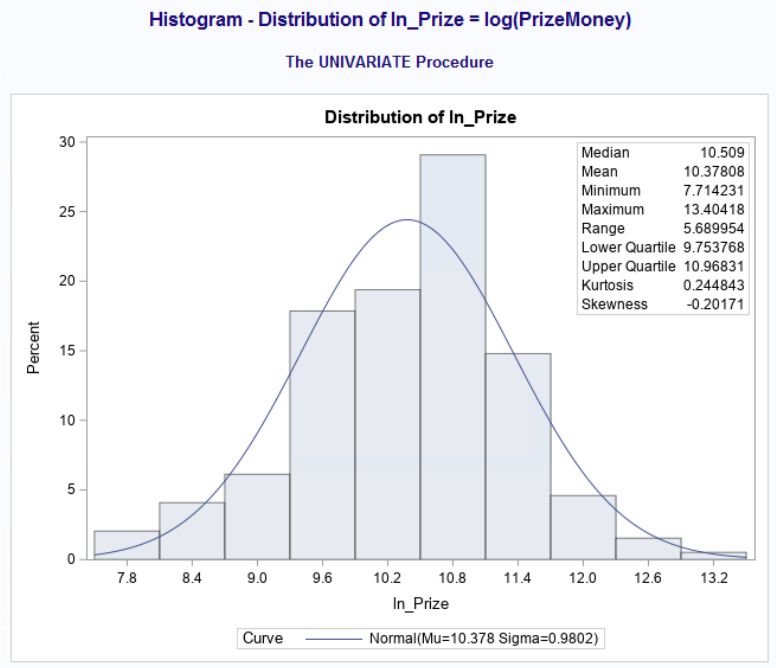
The graph represents the linear relationship between PrizeMoney and other variables (DrivingAccuracy, GIR, PuttingAverage, BirdieConversion, PuttsPerRound).
From the correlation coefficients, we can tell that there is no association between PrizeMoney and DrivingAccuracy with value 0.02468.
GIR has the value as 0.41022 and BirdieConversion has the value as 0.41343. They are moderately positive correlated to PrizeMoney.
PuttingAverage has the value as -0.31305 and PuttsPerRound has the value as -0.11249. They represent mildly negative correlation to PrizeMoney.

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



Among 196 players, the mean of PrizeMoney is $50,891.17. The range starts from $2,240 to $662,771, with a large wide $660,531. The median is $36,644.5 which located in the middle of Q1 and Q3. The mean is greater than the median in this graph, hence the distribution is right and positively skewed and unimodal. A pointed top and lighter tails characterize the distribution. The potential outliers may occur at $640,000.

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.



The distribution is symmetric and normal after log transformation. The median ($10.509) is quite close to the mean ($10.37808) now. The range starts from $7.714231 to $13.40418 with a smaller wide $5.689954. The median ($10.509) is located between Q1 ($9.753768) and Q3 ($10.96831). All in all, the distribution has a normal peak.

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.

Regression Model - All Variables

The REG Procedure
Model: MODEL1
Dependent Variable: ln_Prize

Number of Observations Read	196
Number of Observations Used	196

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	101.43308	20.28662	44.86	<.0001
Error	190	85.92207	0.45222		
Corrected Total	195	187.35515			

Root MSE	0.67247	R-Square	0.5414
Dependent Mean	10.37808	Adj R-Sq	0.5293
Coeff Var	6.47975		

Parameter Estimates

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

Regression Model - Remove DrivingAccuracy

The REG Procedure
Model: MODEL1
Dependent Variable: ln_Prize

Number of Observations Read	196
Number of Observations Used	196

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	101.43115	25.35779	56.37	<.0001
Error	191	85.92400	0.44986		
Corrected Total	195	187.35515			

Root MSE	0.67072	R-Square	0.5414
Dependent Mean	10.37808	Adj R-Sq	0.5318
Coeff Var	6.46284		

Parameter Estimates

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

Regression Model - Remove DrivingAccuracy & PuttingAverage

The REG Procedure
Model: MODEL1
Dependent Variable: ln_Prize

Number of Observations Read	196
Number of Observations Used	196

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	100.17370	33.39123	73.54	<.0001
Error	192	87.18145	0.45407		
Corrected Total	195	187.35515			

Root MSE	0.67385	R-Square	0.5347
Dependent Mean	10.37808	Adj R-Sq	0.5274
Coeff Var	6.49299		

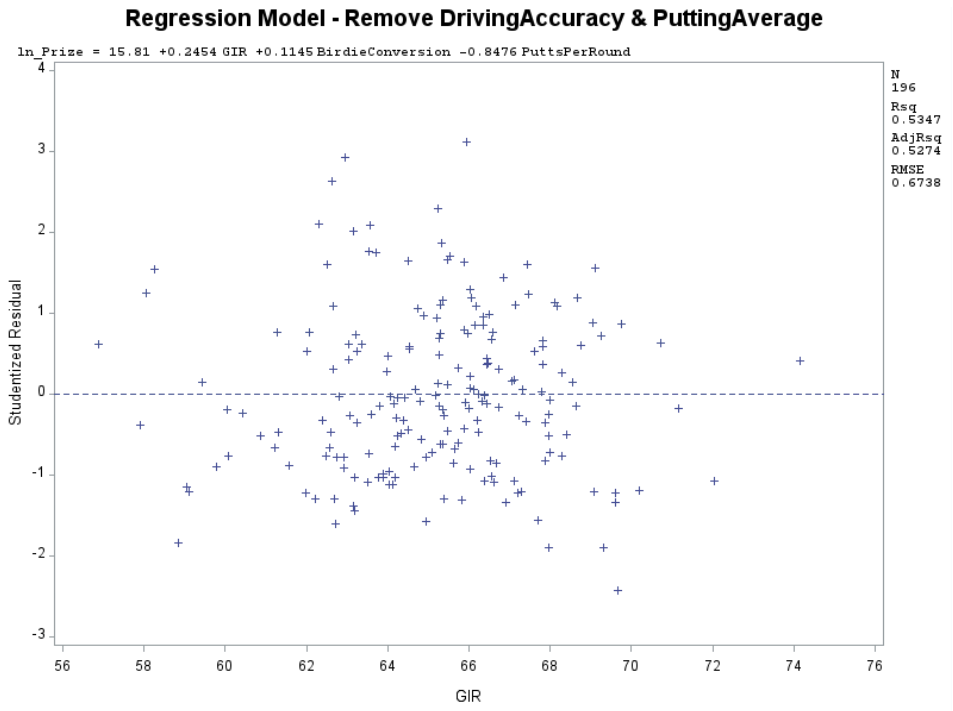
Parameter Estimates

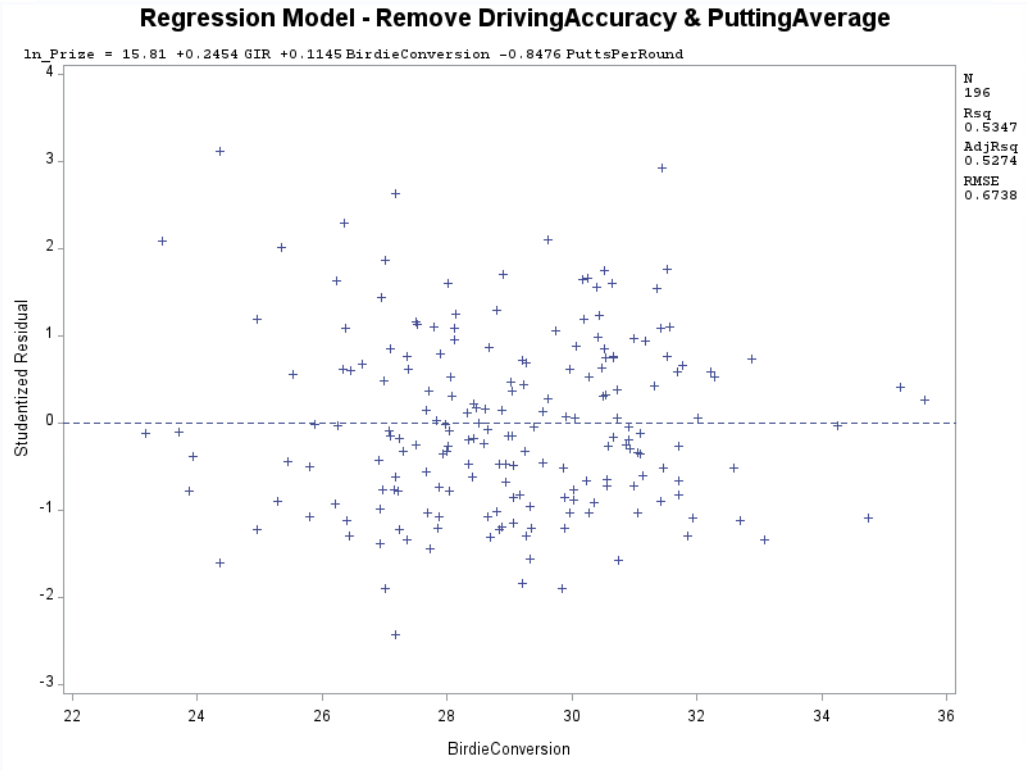
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	15.81016	4.34465	3.64	0.0004
GIR	1	0.24542	0.02160	11.36	<.0001
BirdieConversion	1	0.11454	0.02700	4.24	<.0001
PuttsPerRound	1	-0.84757	0.15377	-5.51	<.0001

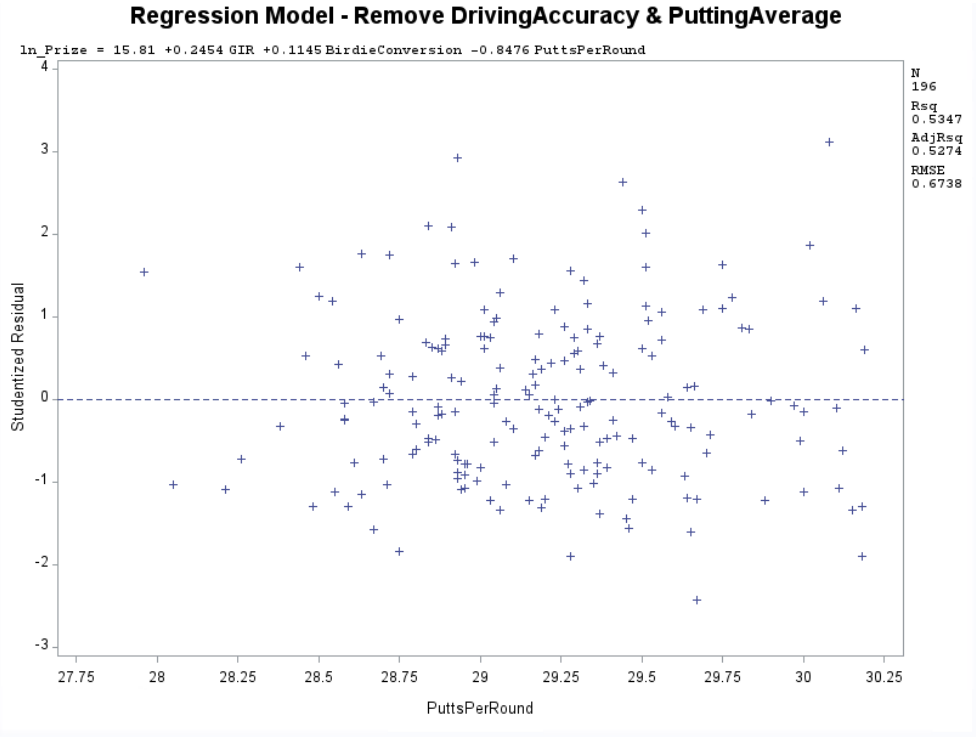
a) If necessary remove the non-significant variables. Remember to remove one variable at a time (variable with largest p-value is removed first) and refit the model, until all variables are significant.

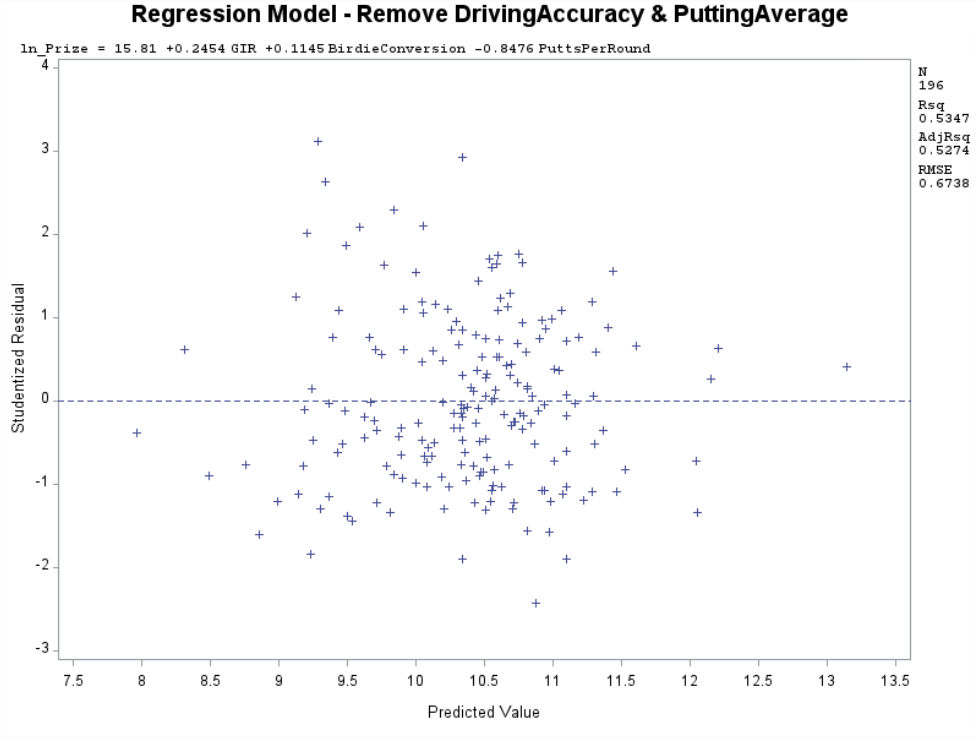
The first graph is a regression model for In_Prize and other all variables (DrivingAccuracy, GIR, PuttingAverage, BirdieConversion and PuttsPerRound). As the p-value for DrivingAccuracy is 0.948, it has the largest p-value among all the others. So the next step is deleting DrivingAccuracy.
The second graph is removing DrivingAccuracy. Then the situation is telling the regression model for In_Prize and other variables (GIR, PuttingAverage, BirdieConversion and PuttsPerRound). The PuttingAverage has the largest p-value as 0.0962 among others. So the next step is removing PuttingAverage.
Finally, after removing PuttingAverage, we can tell from the graph that all the p-value of variables are less than 0.05. In that case GIR, BirdieConversion and PuttsPerRound are significant variables to In_Prize.

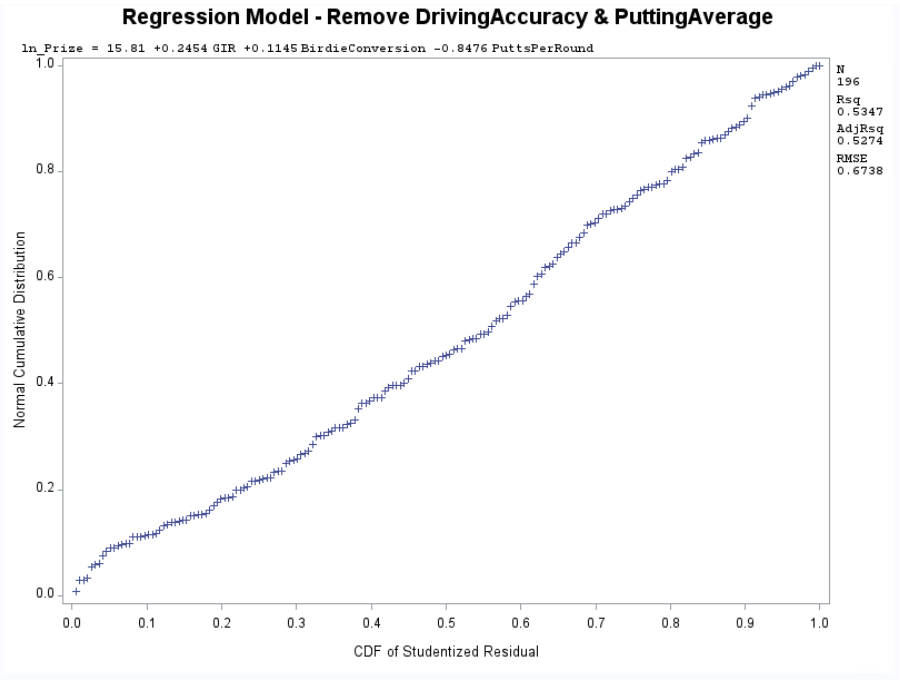
b) Analyze residual plots to check if the regression model is valid for your data. Discuss your analysis.


Regression Model - Remove DrivingAccuracy & PuttingAverage
ln_Prize = 15.81 +0.2454 GIR +0.1145 BirdieConversion -0.8476 PuttsPerRound
N
196
Rsq
0.5347
AdjRsq
0.5274
RMSE
0.6738
Studentized Residual
GIR


Regression Model - Remove DrivingAccuracy & PuttingAverage
ln_Prize = 15.81 +0.2454 GIR +0.1145 BirdieConversion -0.8476 PuttsPerRound
N
196
Rsq
0.5347
AdjRsq
0.5274
RMSE
0.6738
Studentized Residual
BirdieConversion

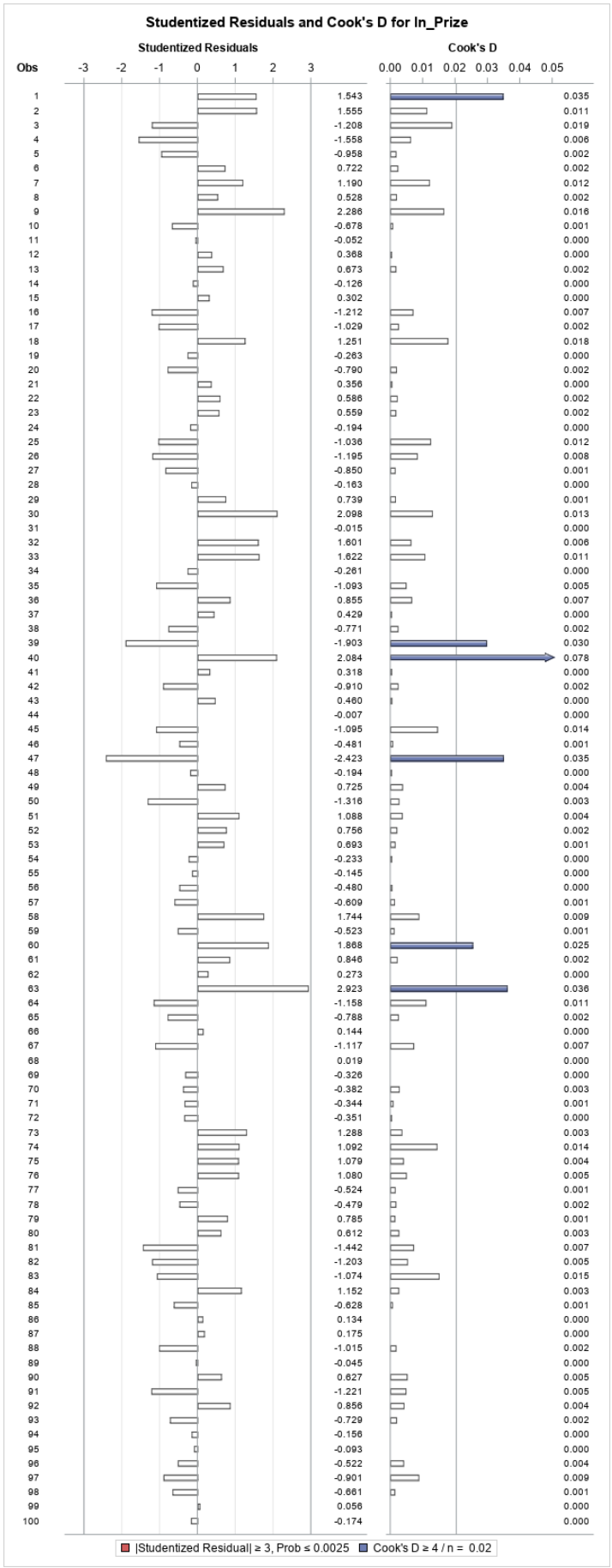

Regression Model - Remove DrivingAccuracy & PuttingAverage
ln_Prize = 15.81 +0.2454 GIR +0.1145 BirdieConversion -0.8476 PuttsPerRound
N
196
Rsq
0.5347
AdjRsq
0.5274
RMSE
0.6738
Studentized Residual
PuttsPerRound
4
3
2
1
0
-1
-2
-3
27.75
28
28.25
28.5
28.75
29
29.25
29.5
29.75
30
30.25

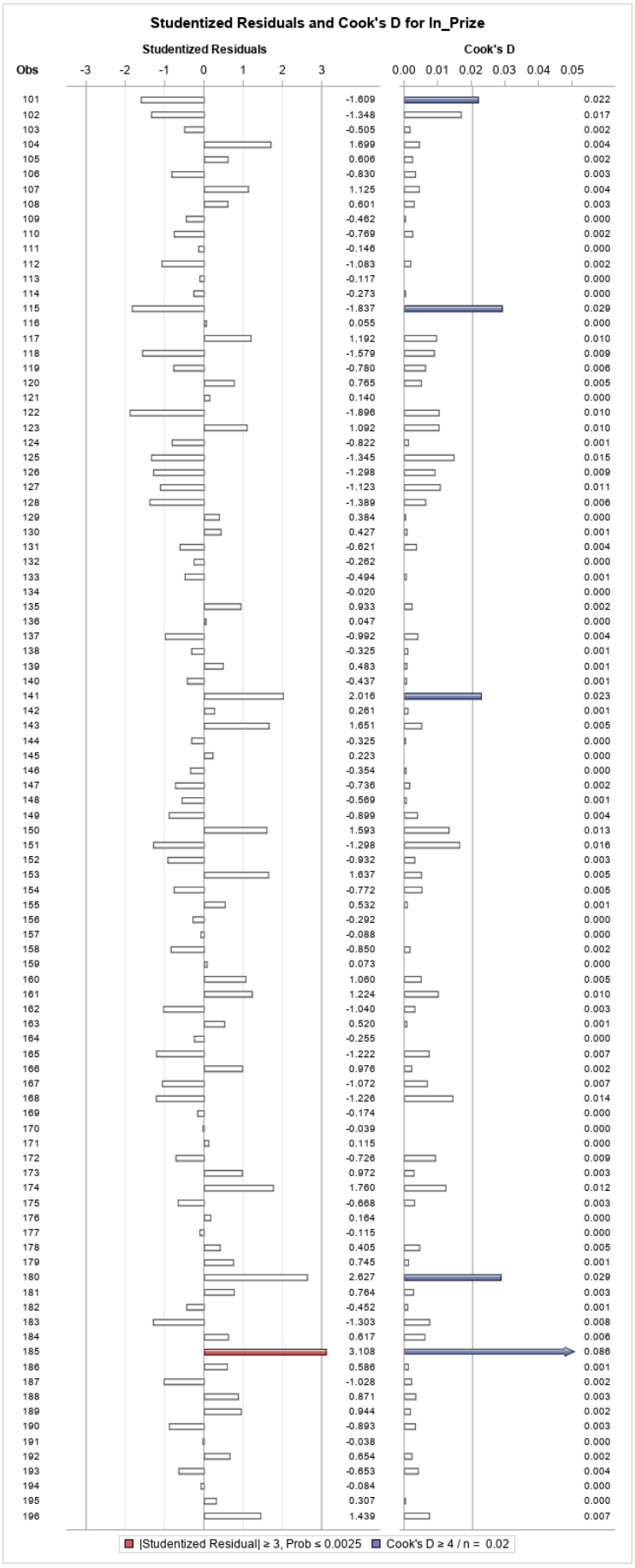

Regression Model - Remove DrivingAccuracy & PuttingAverage
ln_Prize = 15.81 +0.2454 GIR +0.1145 BirdieConversion -0.8476 PuttsPerRound
N
196
Rsq
0.5347
AdjRsq
0.5274
RMSE
0.6738
Studentized Residual
Predicted Value
4
3
2
1
0
-1
-2
-3
7.5
8
8.5
9
9.5
10
10.5
11
11.5
12
12.5
13
13.5



The first three graphs (GIR, BirdiesConversion and PuttsPerRound) are linearly.
The first four graphs (GIR, BirdiesConversion, PuttsPerRound and predicted value) are scattered around zero line by random. They represent independency and constant variance.
The last plot represents a linearly line, it means the model is linearly normal distributed.

c) 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.


Studentized Residuals and Cook's D for ln_Prize
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.00 0.01 0.02 0.03 0.04 0.05
1 1.543 0.035
2 1.555 0.011
3 -1.208 0.019
4 -1.558 0.006
5 -0.958 0.002
6 0.722 0.002
7 1.190 0.012
8 0.528 0.002
9 2.286 0.016
10 -0.678 0.001
11 -0.052 0.000
12 0.368 0.000
13 0.673 0.002
14 -0.126 0.000
15 0.302 0.000
16 -1.212 0.007
17 -1.029 0.002
18 1.251 0.018
19 -0.263 0.000
20 -0.790 0.002
21 0.356 0.000
22 0.586 0.002
23 0.559 0.002
24 -0.194 0.000
25 -1.036 0.012
26 -1.195 0.008
27 -0.850 0.001
28 -0.163 0.000
29 0.739 0.001
30 2.098 0.013
31 -0.015 0.000
32 1.601 0.006
33 1.622 0.011
34 -0.261 0.000
35 -1.093 0.005
36 0.855 0.007
37 0.429 0.000
38 -0.771 0.002
39 -1.903 0.030
40 2.084 0.078
41 0.318 0.000
42 -0.910 0.002
43 0.460 0.000
44 -0.007 0.000
45 -1.095 0.014
46 -0.481 0.001
47 -2.423 0.035
48 -0.194 0.000
49 0.725 0.004
50 -1.316 0.003
51 1.088 0.004
52 0.756 0.002
53 0.693 0.001
54 -0.233 0.000
55 -0.145 0.000
56 -0.480 0.000
57 -0.609 0.001
58 1.744 0.009
59 -0.523 0.001
60 1.868 0.025
61 0.846 0.002
62 0.273 0.000
63 2.923 0.036
64 -1.158 0.011
65 -0.788 0.002
66 0.144 0.000
67 -1.117 0.007
68 0.019 0.000
69 -0.326 0.000
70 -0.382 0.003
71 -0.344 0.001
72 -0.351 0.000
73 1.288 0.003
74 1.092 0.014
75 1.079 0.004
76 1.080 0.005
77 -0.524 0.001
78 -0.479 0.002
79 0.785 0.001
80 0.612 0.003
81 -1.442 0.007
82 -1.203 0.005
83 -1.074 0.015
84 1.152 0.003
85 -0.628 0.001
86 0.134 0.000
87 0.175 0.000
88 -1.015 0.002
89 -0.045 0.000
90 0.627 0.005
91 -1.221 0.005
92 0.856 0.004
93 -0.729 0.002
94 -0.156 0.000
95 -0.093 0.000
96 -0.522 0.004
97 -0.901 0.009
98 -0.661 0.001
99 0.056 0.000
100 -0.174 0.000
|Studentized Residual| ≥ 3, Prob ≤ 0.0025
Cook's D ≥ 4 / n = 0.02


Studentized Residuals and Cook's D for ln_Prize
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.00 0.01 0.02 0.03 0.04 0.05
101 -1.609 0.022
102 -1.348 0.017
103 -0.505 0.002
104 1.699 0.004
105 0.606 0.002
106 -0.830 0.003
107 1.125 0.004
108 0.601 0.003
109 -0.462 0.000
110 -0.769 0.002
111 -0.146 0.000
112 -1.083 0.002
113 -0.117 0.000
114 -0.273 0.000
115 -1.837 0.029
116 0.055 0.000
117 1.192 0.010
118 -1.579 0.009
119 -0.780 0.006
120 0.765 0.005
121 0.140 0.000
122 -1.896 0.010
123 1.092 0.010
124 -0.822 0.001
125 -1.345 0.015
126 -1.298 0.009
127 -1.123 0.011
128 -1.389 0.006
129 0.384 0.000
130 0.427 0.001
131 -0.621 0.004
132 -0.262 0.000
133 -0.494 0.001
134 -0.020 0.000
135 0.933 0.002
136 0.047 0.000
137 -0.992 0.004
138 -0.325 0.001
139 0.483 0.001
140 -0.437 0.001
141 2.016 0.023
142 0.261 0.001
143 1.651 0.005
144 -0.325 0.000
145 0.223 0.000
146 -0.354 0.000
147 -0.736 0.002
148 -0.569 0.001
149 -0.899 0.004
150 1.593 0.013
151 -1.298 0.016
152 -0.932 0.003
153 1.637 0.005
154 -0.772 0.005
155 0.532 0.001
156 -0.292 0.000
157 -0.088 0.000
158 -0.850 0.002
159 0.073 0.000
160 1.060 0.005
161 1.224 0.010
162 -1.040 0.003
163 0.520 0.001
164 -0.255 0.000
165 -1.222 0.007
166 0.976 0.002
167 -1.072 0.007
168 -1.226 0.014
169 -0.174 0.000
170 -0.039 0.000
171 0.115 0.000
172 -0.726 0.009
173 0.972 0.003
174 1.760 0.012
175 -0.668 0.003
176 0.164 0.000
177 -0.115 0.000
178 0.405 0.005
179 0.745 0.001
180 2.627 0.029
181 0.764 0.003
182 -0.452 0.001
183 -1.303 0.008
184 0.617 0.006
185 3.108 0.086
186 0.586 0.001
187 -1.028 0.002
188 0.871 0.003
189 0.944 0.002
190 -0.893 0.003
191 -0.038 0.000
192 0.654 0.002
193 -0.653 0.004
194 -0.084 0.000
195 0.307 0.000
196 1.439 0.007
|Studentized Residual| ≥ 3, Prob ≤ 0.0025
Cook's D ≥ 4 / n = 0.02

There is only one outlier at obs 185 marked as red color. Besides, the influential points are marked in blue. Firstly, let's say obs 185 is marked with both outlier and influential point. We will remove it. Then we retry it and see the changes on R-Square and adj R-Square and whether it helps to improve the result until nothing will be changed at the end. All in all, I tried removing 17 obs when R-Square and adj R-Square are the largest.

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

The REG Procedure
Model: MODEL1
Dependent Variable: ln_Prize

Number of Observations Read	179
Number of Observations Used	179

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	117.24024	39.08008	128.46	<.0001
Error	175	53.23855	0.30422		
Corrected Total	178	170.47879			

Root MSE	0.55156	R-Square	0.6877
Dependent Mean	10.35579	Adj R-Sq	0.6824
Coeff Var	5.32612		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.62487	3.78413	3.07	0.0025
GIR	1	0.29826	0.01935	15.41	<.0001
BirdieConversion	1	0.13252	0.02362	5.61	<.0001
PuttsPerRound	1	-0.84227	0.13239	-6.36	<.0001

Final model equation:
$\ln\_Prize = 11.62487 + 0.29826 * GIR + 0.13252 * BirdiesConversion - 0.84227 * PuttsPerRound + e$

Null hypothesis:

$H_o: \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

Alternative hypothesis:

H_a: At least one coefficient $\beta_j \neq 0$

Test statistic:

This model is the final model because it has the highest R-Square (0.6877) and adj R-Square (0.6824). F = 128.46 and with p-value less than 0.001 (at alpha=0.05). The null hypothesis of no

association between PrizeMoney and other variables (GIR, BirdiesConversion and PuttsPerRound) is rejected. At least one x-variable has a significant effect on changes in PrizeMoney. F-test gives strong support to the fitted model.

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?

GIR has positive association to average prize money. As all the other variables keep constantly, every 1% increase of GIR will make average prize money increases 34.99%.

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

```
TITLE "Analysis of pgatour2006";

PROC IMPORT datafile="C:\Users\XLIU115\Desktop\Assignment4\pgatour2006.csv"
out=PGATour replace;
delimiter=',';
getnames=yes;
RUN;

PROC PRINT data=PGATour;
RUN;

/*2A*/
PROC SGSCATTER;
title "Scatterplot Matrix - PrizeMoney and Other Variables";
matrix PrizeMoney DrivingAccuracy GIR PuttingAverage BirdieConversion
PuttsPerRound;
RUN;

/*2B*/
TITLE "Histogram - Distribution of PrizeMoney";

PROC UNIVARIATE normal;
var PrizeMoney;
histogram / normal (mu=est sigma=est);
inset median mean min max range Q1 Q3 kurtosis skewness /pos = ne;
RUN;

/*2C*/
DATA PGATour;
set PGATour;
ln_Prize = log(PrizeMoney);

PROC PRINT;
RUN;

TITLE "Histogram - Distribution of ln_Prize = log(PrizeMoney)";
PROC UNIVARIATE normal;
var ln_Prize;
histogram / normal (mu=est sigma=est);
inset median mean min max range Q1 Q3 kurtosis skewness /pos = ne;
RUN;

/*2D*/
```

```
/*Regression Model - All Variables*/
PROC REG;
title "Regression Model - All Variables";
model ln_Prize = DrivingAccuracy GIR PuttingAverage BirdieConversion
PuttsPerRound;
RUN;

PROC REG corr;
model ln_Prize = DrivingAccuracy GIR PuttingAverage BirdieConversion
PuttsPerRound;
RUN;


/*Regression Model - Remove DrivingAccuracy*/
PROC REG;
title "Regression Model - Remove DrivingAccuracy";
model ln_Prize = GIR PuttingAverage BirdieConversion PuttsPerRound;
RUN;

PROC REG corr;
model ln_Prize = GIR PuttingAverage BirdieConversion PuttsPerRound;
RUN;


/*Regression Model - Remove DrivingAccuracy & PuttingAverage*/
PROC REG;
title "Regression Model - Remove DrivingAccuracy & PuttingAverage";
model ln_Prize = GIR BirdieConversion PuttsPerRound /influence r;
plot student.*(GIR BirdieConversion PuttsPerRound predicted.);
plot npp.*student.;
RUN;

PROC REG corr;
model ln_Prize = GIR BirdieConversion PuttsPerRound;
RUN;


DATA PGATour2;
set PGATour2;
if _n_ in (138) then delete;
RUN;

PROC REG;
title "Regression Model - Remove DrivingAccuracy & PuttingAverage New";
model ln_Prize = GIR BirdieConversion PuttsPerRound /influence r;
plot student.*(GIR BirdieConversion PuttsPerRound predicted.);
plot npp.*student.;
RUN;
```

Problem 3 [20 pts] – ONLY for GRADUATES

Answer the following questions:

1. Why is Adj-R^2 a better indicator to use than the R^2? Explain. (1 point)

   Adj-R^2 is the better model when comparing models that have a different amount of variables. R^2 always increases when the number of variables increases. Even if adding a useless variable to model, R^2will still increase. We should always compare models with different number of independent variables with Adj-R^2. Adj-R^2 only increases if the new variable improves the model more than would be expected by chance.

2. Which transformations will you use for 'Loan Amount'? List all that applies. (1 point)

   1) Log(Y) (only if Y > 0)

   2) Sqrt(Y) (only if Y >= 0)

   3) Square Y =Y2

   4) Cubic Y = Y3

   5) Inverse Y = 1/Y (only for Y ≠0)

3. At which stages should you check for assumptions? (2 points)

   Fit the full model (i.e. with all the predictors) and check residuals/assumptions.

   Final Model (i.e. final fitted model) and check residuals/assumptions.

4. At which stages should you check for collinearity? (2 points)

   At Data Exploration Stage.

   At Analysis State – Verify & Confirm.

5. At which stages should you check for outliers? (2 points)

   At Data Exploration Stage.

At Analysis Stage.

6. At which stages should you check for influential points? (2 points)

   At Data Exploration Stage.

   At Analysis Stage.

7. "Outliers and influential points are the same" – explain why this statement is true/false. (2 points)

   False.

   Outliers are the data points those diverge by good margin from the overall pattern. It can have an extreme X or Y values or both compared to other values.

   Influential point is an outlier that impacts the slope of the regression line. To test the influence of an outlier is to compute the regression equation with and without the outlier.

8. When should I ignore outliers and influential points? Explain. (2 points)

   Removing an observation.

   Remove the outlier obs# and run the model. Check the adj-R2, residual plots and p-values of the predictors. See if they improve.

   Remove the influential point that got flagged by almost all indicators obs. Check the adj-R2, residual plots and p-values of the predictors. See if they improve. If it doesn't, keep it as part of observations.

   Rerun until check adj-R2, goodness of fit test, residuals and p-values of all predictors. If Adj-R2 get improved, f-value is high, and p-value associated with f-statistic is less than 0.05, then overall goodness of fit test shows that at least one predictor is significantly associated with Y.

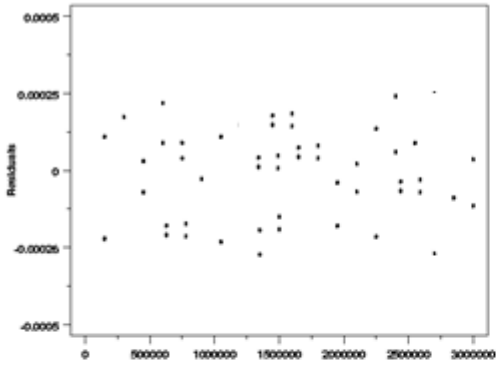
   Then we can ignore outliers and influential points.

9. When encountered with collinearity issue, one option is to drop one of the collinear variable.

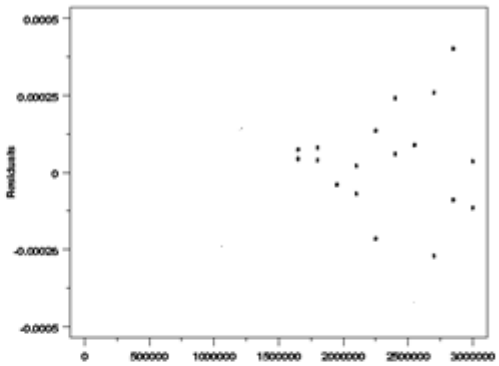
   What other option(s) do I have other than dropping one variable? (2 points)

   Instead of doing nothing, the other options are
   Create 2 different models, one with one of the collinear variables and the 2nd model with the other collinear variable
   Reuse the variable by creating another variable
   Create interaction terms
   Centering the collinear variables

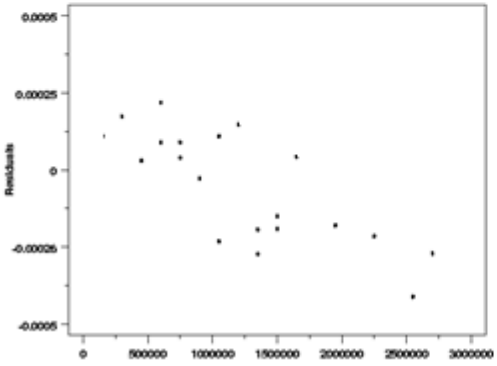
10. Based on the residual plots shown for X1, X2, X3 and predicted Y variables, indicate which of these you will need to transform. Explain. (4 points)



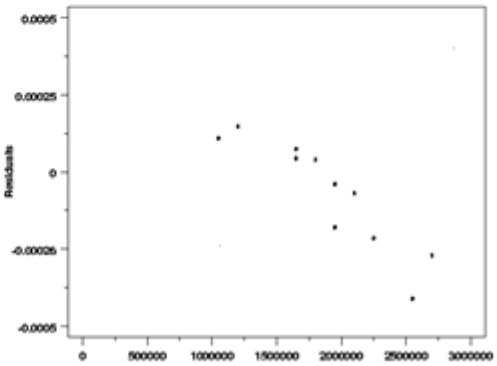
X1



X2



X3



Predicted Y

X2, X3 and Predicted Y need to transformation. Because The residual plots show points that are NOT randomly scattered or the normal probability plot shows "S" shape. In that case we need to apply transformation on the response variable Y to stabilize the variance.