Question 4

Part a

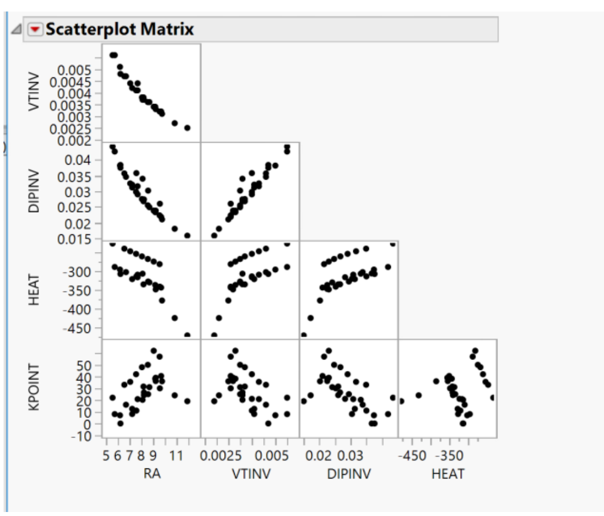
The model cannot be claimed as valid.
This is confirmed from the residual plots and correlation plots in JMP along with fitting a linear regression model. Only VTINV is having significant p-value.
The residual plots implies the error terms are non-random. Transformations could be tried.
There is a correlation between some predictors from scatterplot matrix. Multi-collinearity should be checked for.

Summary of Fit

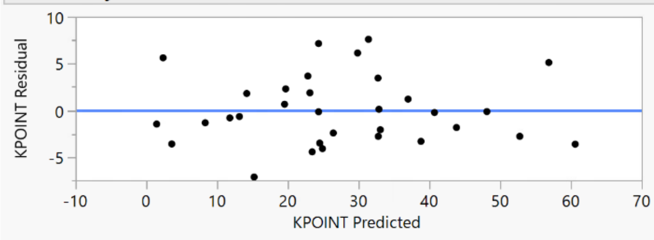
RSquare	0.944562
RSquare Adj	0.936349
Root Mean Square Error	3.919278
Mean of Response	27.35
Observations (or Sum Wgts)	32

Parameter Estimates

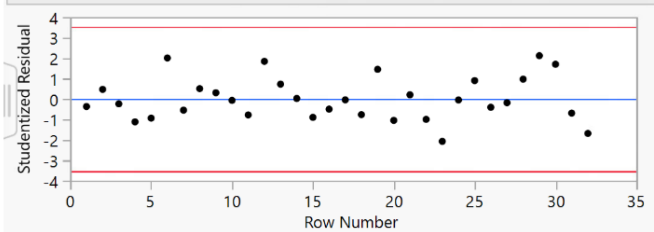
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	70.312696	33.67576	2.09	0.0464*
RA	10.472754	2.418209	4.33	0.0002*
VTINV	9038.1905	4409.411	2.05	0.0502
DIPINV	-1826.242	376.5281	-4.85	<.0001*
HEAT	0.3550077	0.021764	16.31	<.0001*



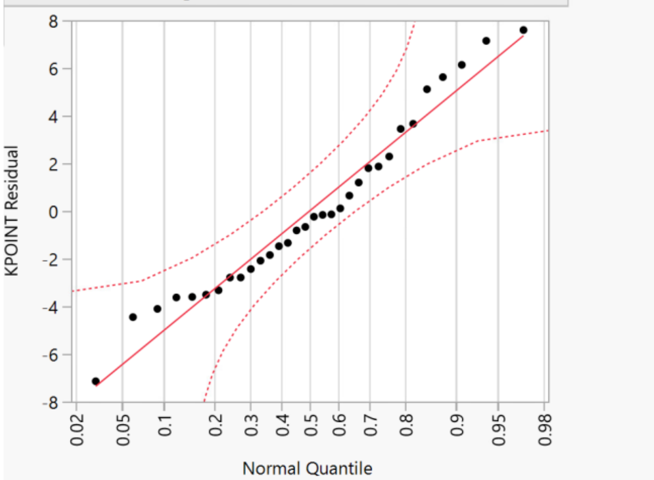
Residual by Predicted Plot



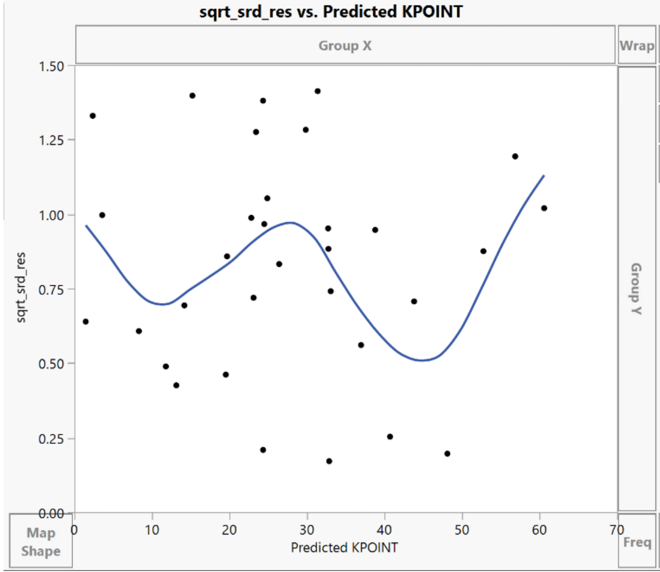
Studentized Residuals

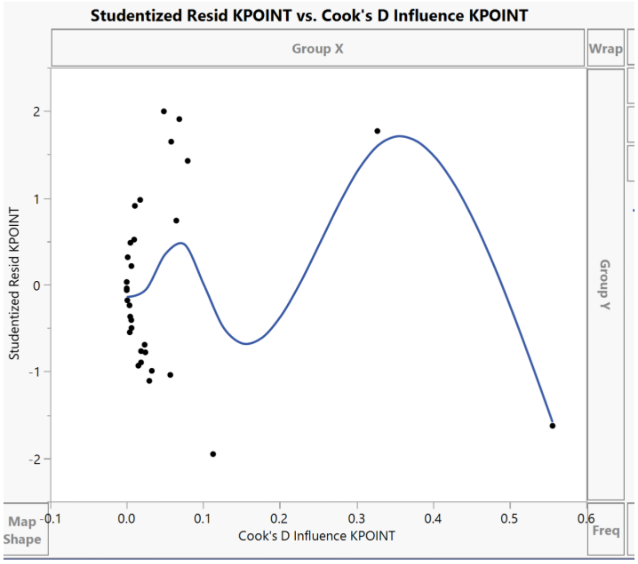


Residual Normal Quantile Plot



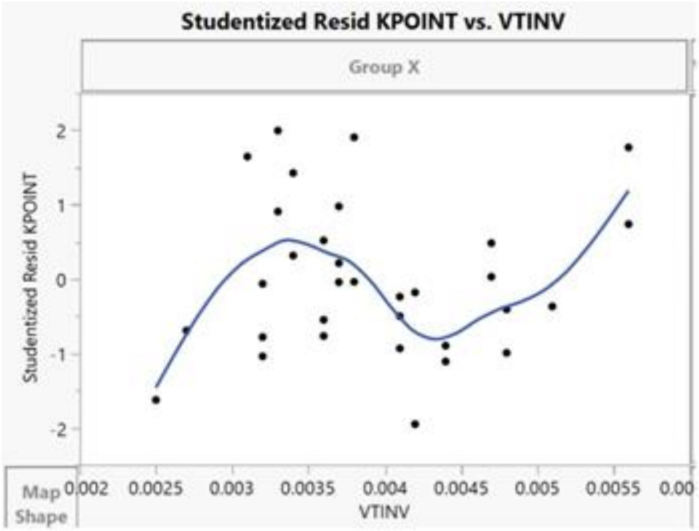
sqrt_srd_res vs. Predicted KPOINT

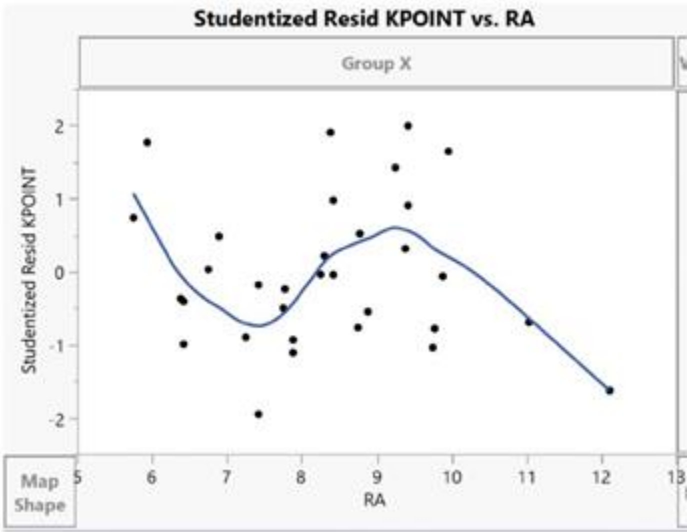




Part- B

Plots of the dependent variable with RA and VTINV is not random. This is confirmed from correlation plot matrix. On fitting a linear model, the standardized residual vs RA / VTINV would be curved too.
So, this implies, we should include some squared terms of both the variables to fit the pattern better.



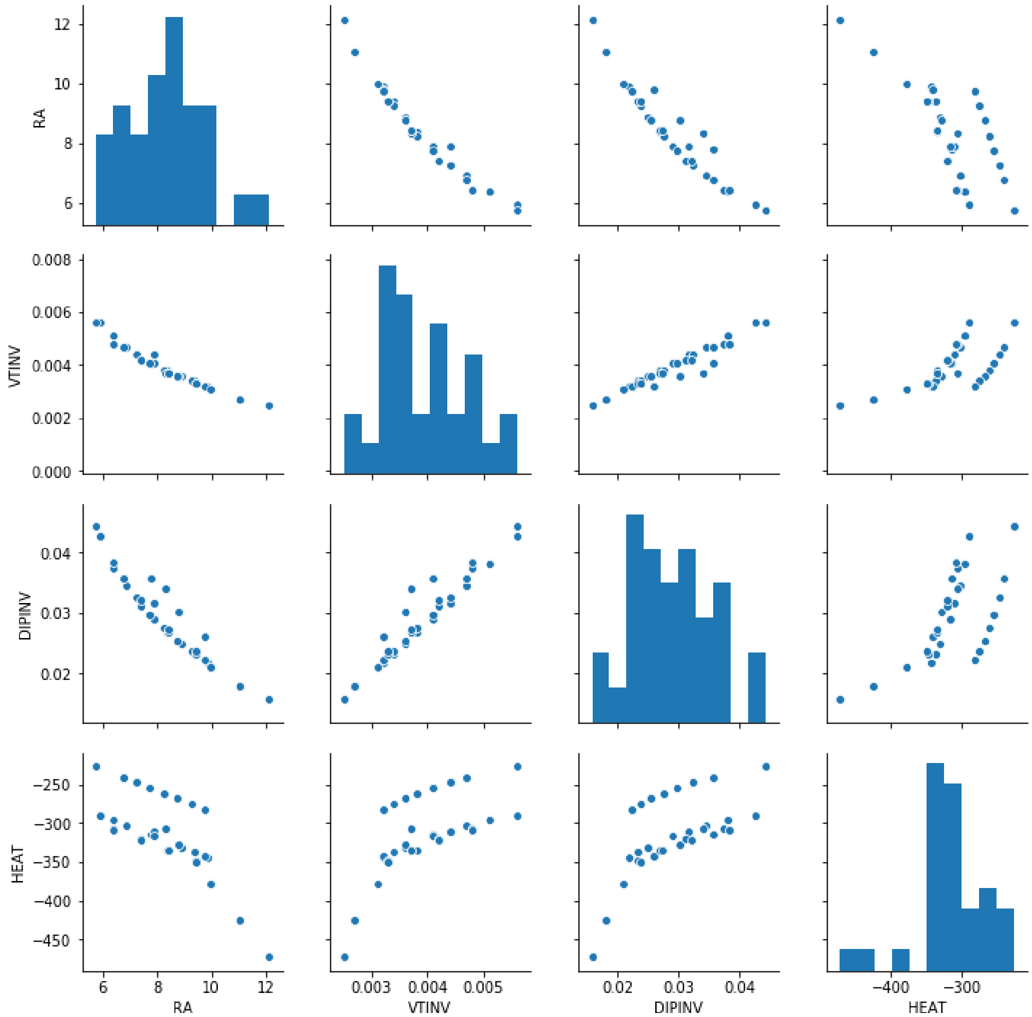


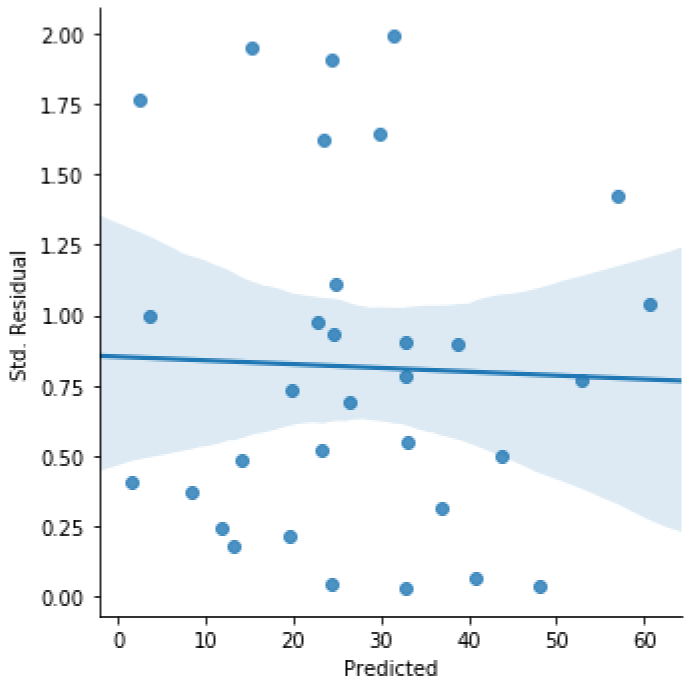
Part- C

The four criteria proposed by Jalali-Heravi and Knouz is not the best to compare models. It does talk about Error behavior. In the BLUE assumptions of linear regression, three are- independence of error terms, normal distribution of residuals and constant variance. The best model should satisfy these assumptions and then based on metrics like R-square value or MSE, best model can be selected.

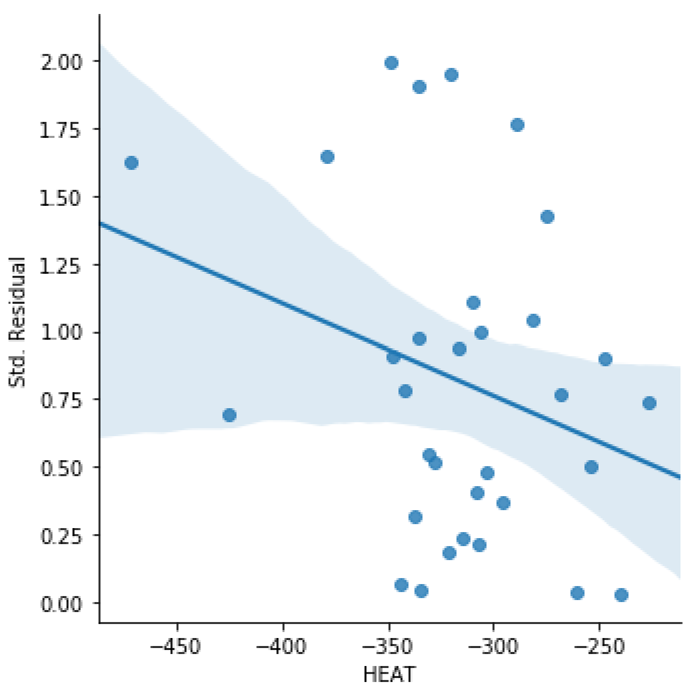
Python solution for the same:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 KPOINT   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.936
Method:                 Least Squares   F-statistic:                     115.0
Date:                Fri, 08 Mar 2019   Prob (F-statistic):           1.51e-16
Time:                        21:49:48   Log-Likelihood:                -86.397
No. Observations:                  32   AIC:                             182.8
Df Residuals:                      27   BIC:                             190.1
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     70.3127     33.676      2.088      0.046       1.216     139.410
RA            10.4728      2.418      4.331      0.000       5.511      15.435
VTINV       9038.1905   4409.411      2.050      0.050      -9.173    1.81e+04
DIPINV     -1826.2421    376.528     -4.850      0.000   -2598.814   -1053.670
HEAT           0.3550      0.022     16.312      0.000       0.310       0.400
==============================================================================
Omnibus:                        1.526   Durbin-Watson:                   1.807
Prob(Omnibus):                  0.466   Jarque-Bera (JB):                1.434
Skew:                           0.456   Prob(JB):                        0.488
Kurtosis:                       2.508   Cond. No.                     2.04e+06
==============================================================================
```

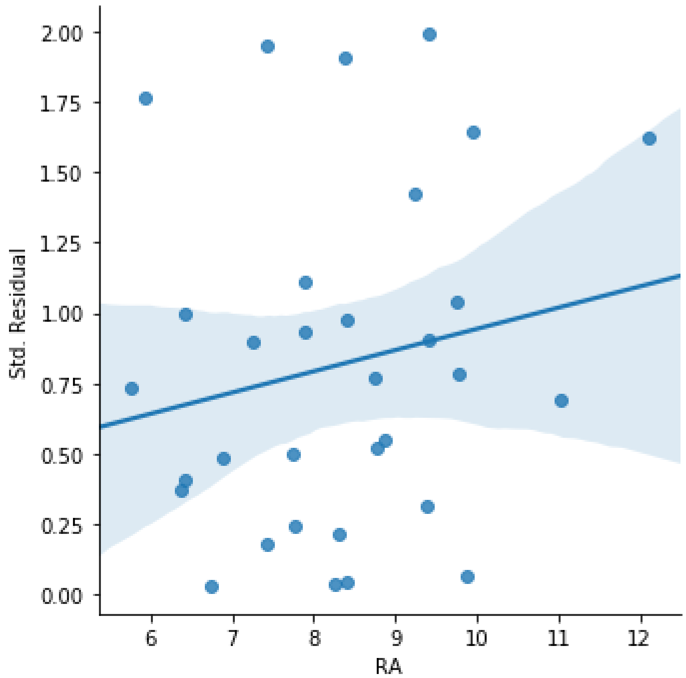

RA
VTINV
DIPINV
HEAT

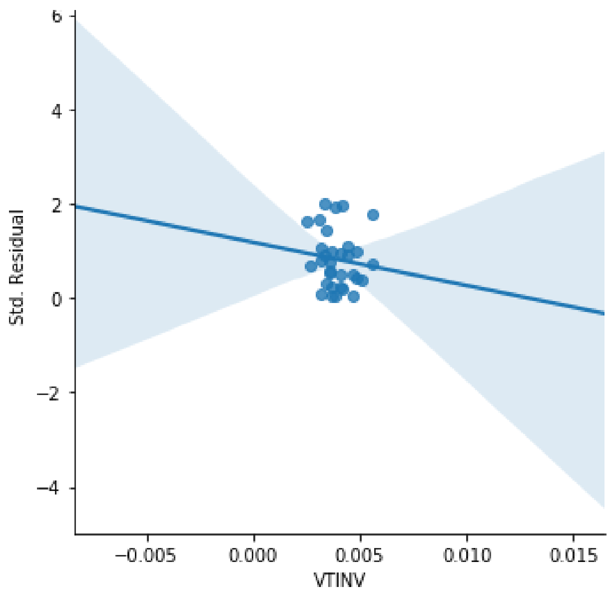


Std. Residual
Predicted
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0
10
20
30
40
50
60

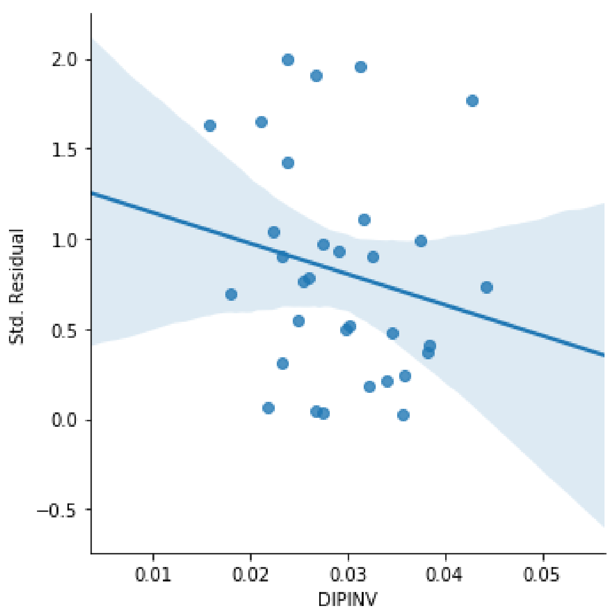


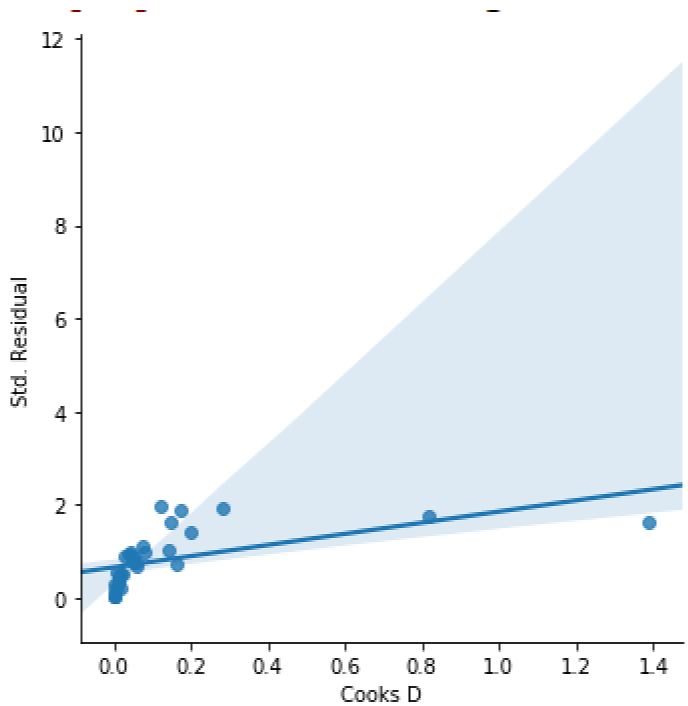
Std. Residual
HEAT
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
−450
−400
−350
−300
−250


Std. Residual
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
6
7
8
9
10
11
12
RA




```
In [48]: sns.lmplot('DIPINV','Std. Residual',data=df)
Out[48]: <seaborn.axisgrid.FacetGrid at 0x167486ff748>
```






```
import math
import pandas as pd
import numpy as np
import statsmodels.formula.api as sm
from statsmodels.sandbox.regression.predstd import wls_prediction_std
import matplotlib.pyplot as plt
import seaborn as sns
path = "C:/Users/haris/Downloads/pgatour2006.csv"
df = pd.read_csv(path, sep = ",")
df['PrizeMoney']=df['PrizeMoney'].transform(np.log)
print(df[0:10], '\n')
model = sm.ols('PrizeMoney ~ DrivingAccuracy + GIR
+PuttingAverage+BirdieConversion+SandSaves+Scrambling+BounceBack+PuttsPerRound',
data=df)
results = model.fit()
print(results.summary(), "\n")
# Save Predictions
pred = results.fittedvalues
# Save Residuals
resid = results.resid
# Calculate Hii
results.HC2_se
het = results.het_scale # het = r^2 / (1-Hii)
h = 1.0 - (1.0/het)resid*2 # h = Hii
# Calculate Standardized Residuals
std_resid = np.sqrt(het)/math.sqrt(results.mse_resid)
std_resid = pd.Series(std_resid) # Move into Pandas Series
# Correct for negative signs
```

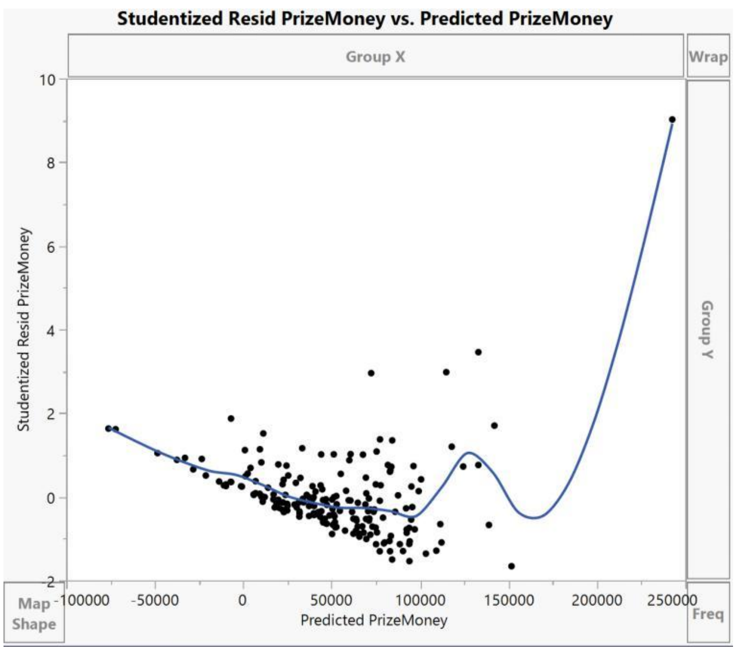
```
for i in range(df.shape[0]):
    if resid[i] < 0:
        std_resid[i] = -std_resid[i]
# Calculate Cook's D
D = ((std_resid*2)*h)/(2(1.0-h))
df2 = pd.concat([pred, resid, std_resid, D, h], axis=1, \
keys=["Predicted", "Residual", "Std. Residual", \
"Cooks D", "H-Hat"])
df = df.join(df2)
print(df, "\n")
print("Rule of Thumb for Cook's D: ", 4/df.shape[0])
# Plot Predicted vs Observed
fig, ax = plt.subplots(figsize=(8,6))
ax.set_title("Predicted vs Observed", fontweight="bold", fontsize="14")
ax.set_xlabel("Size", fontweight="bold", fontsize="12")
ax.set_ylabel("Price", fontweight="bold", fontsize="12")
ax.plot(df['Size'], df['Price'], 'o', label="Data")
sns.pairplot(df, vars=[ 'RA','VTINV','DIPINV','HEAT'])
legend = ax.legend(loc='best')
plt.savefig("../graphs/Pred_vs_Obs.pdf")
plt.show()
fig, ax = plt.subplots(figsize=(8,6))
ax.set_title("Predicted vs Standardized Residuals", \
fontweight="bold", fontsize="14")
ax.set_xlabel("Price", fontweight="bold", fontsize="12")
ax.set_ylabel("Std. Residual", fontweight="bold", fontsize="12")
sns.lmplot('Predicted','Std. Residual',data=df)
ax.axhline(y=0, linewidth=2, color='b', linestyle='--')
ax.axhline(y=2, linewidth=2, color='r', linestyle='-')
ax.axhline(y=-2, linewidth=2, color='r', linestyle='-')
plt.savefig("../graphs/Pred_vs_SResid.pdf")
plt.show()
fig, ax = plt.subplots(figsize=(8,6))
ax.set_title("Predicted vs Cook's D", \
fontweight="bold", fontsize="14")
ax.set_xlabel("Price", fontweight="bold", fontsize="12")
ax.set_ylabel("Cook's D", fontweight="bold", fontsize="12")
sns.scatterplot(data=df['Std. Residual'])
ax.axhline(y=4/df.shape[0], linewidth=2, color='r', linestyle='-')
plt.savefig("../graphs/Pred_vs_CooksD.pdf")
plt.show()
 # Plot Cook's D vs Std. Residuals
fig, ax = plt.subplots(figsize=(8,6))
ax.set_title("Cook's D vs Std. Residuals", \
fontweight="bold", fontsize="14")
ax.set_xlabel("Cook's D", fontweight="bold", fontsize="12")
ax.set_ylabel("Std. Residual", fontweight="bold", fontsize="12")
sns.lmplot('Cooks D','Std. Residual', data=df)
ax.axhline(y=0, linewidth=2, color='b', linestyle='--')
ax.axhline(y=2, linewidth=2, color='r', linestyle='-')
```

```
ax.axhline(y=-2, linewidth=2, color='r', linestyle='-')
ax.axvline(x=0.0816, linewidth=2, color='g', linestyle='-')
plt.savefig("../graphs/CooksD_SResid.pdf")
plt.show()
```

Question 5

Part – A

Initially, simple regression model was tried without any transformation. The error plot exhibits non-constant variance. R square is ~0.4 with few insignificant variables



Summary of Fit

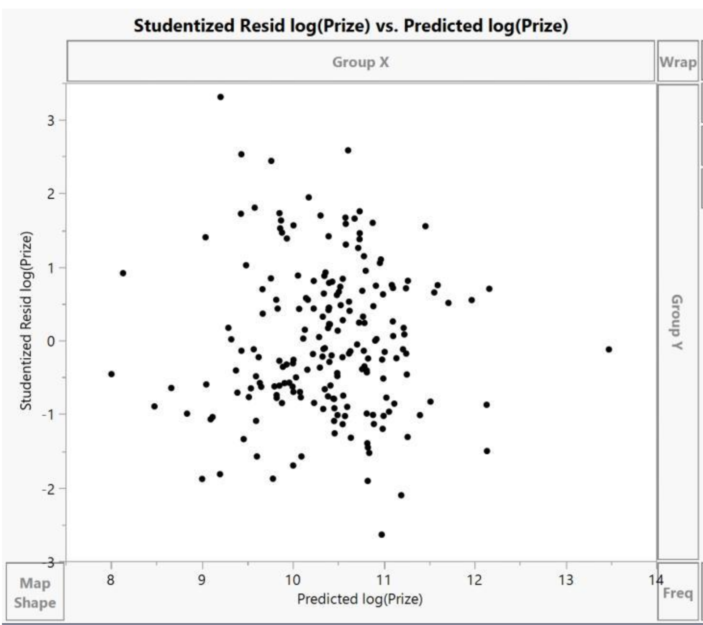
RSquare	0.40639
RSquare Adj	0.384287
Root Mean Square Error	50142.97
Mean of Response	50891.17
Observations (or Sum Wgts)	196

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	3.2361e+11	4.623e+10	18.3866
Error	188	4.7269e+11	2.5143e+9	**Prob > F**
C. Total	195	7.963e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1165233	587382.9	-1.98	0.0487*
DrivingAccuracy	-1835.83	889.1612	-2.06	0.0403*
GIR	9671.3343	3309.355	2.92	0.0039*
PuttingAverage	-47435.3	521566.4	-0.09	0.9276
BirdieConversion	10426.032	3049.642	3.42	0.0008*
SandSaves	1182.0577	744.818	1.59	0.1142
Scrambling	4741.2582	2400.818	1.97	0.0497*
PuttsPerRound	5267.517	35765.74	0.15	0.8831



Trying out log transformation for the response variable generates studentized residual plot as shown above. R square improves to 0.55

Summary of Fit

RSquare	0.557709
RSquare Adj	0.54124
Root Mean Square Error	0.663908
Mean of Response	10.37808
Observations (or Sum Wgts)	196

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	104.48959	14.9271	33.8656
Error	188	82.86555	0.4408	**Prob > F**
C. Total	195	187.35515		<.0001*

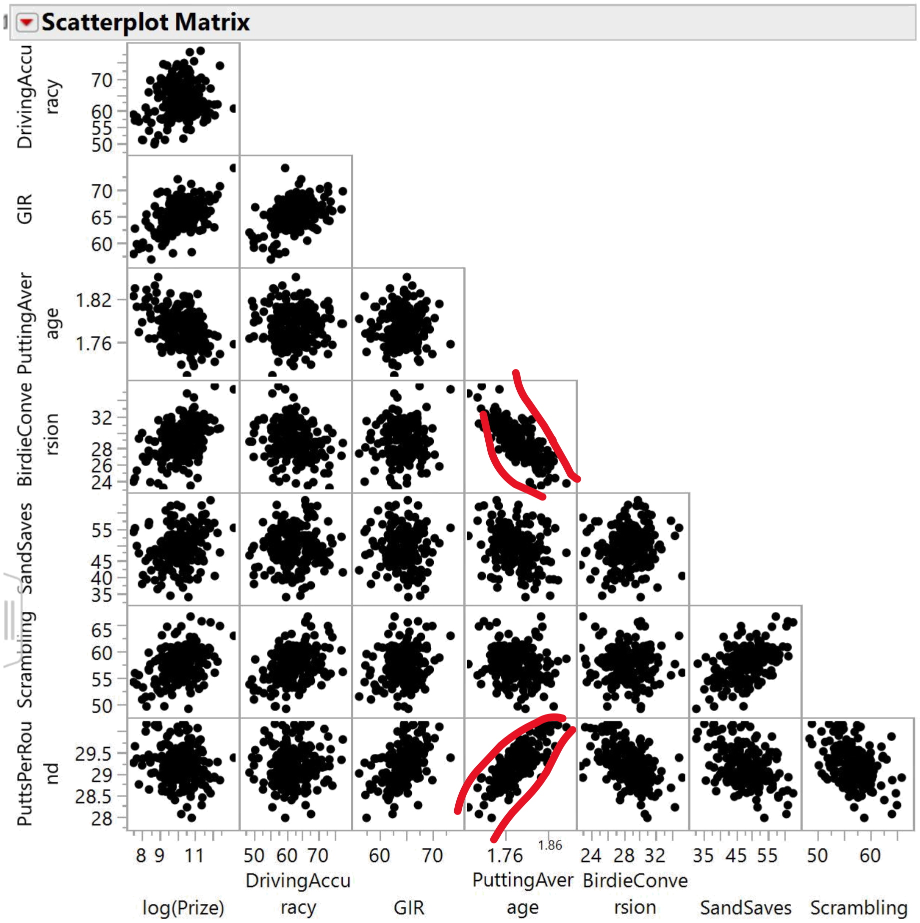
Parameter Estimates

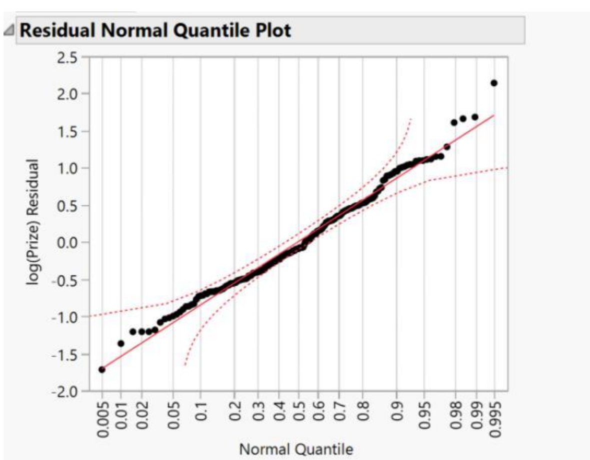
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.1943003	7.777129	0.02	0.9801
DrivingAccuracy	-0.00353	0.011773	-0.30	0.7646
GIR	0.1993109	0.043817	4.55	<.0001*
PuttingAverage	-0.466304	6.905698	-0.07	0.9462
BirdieConversion	0.1573409	0.040378	3.90	0.0001*
SandSaves	0.0151744	0.009862	1.54	0.1256
Scrambling	0.0515137	0.031788	1.62	0.1068
PuttsPerRound	-0.343131	0.473549	-0.72	0.4696

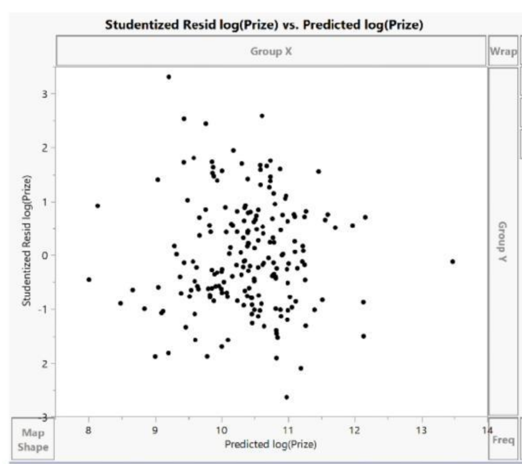
Part – B

The scatterplot matrix tells that there is no direct linear relationship between predictor and response.

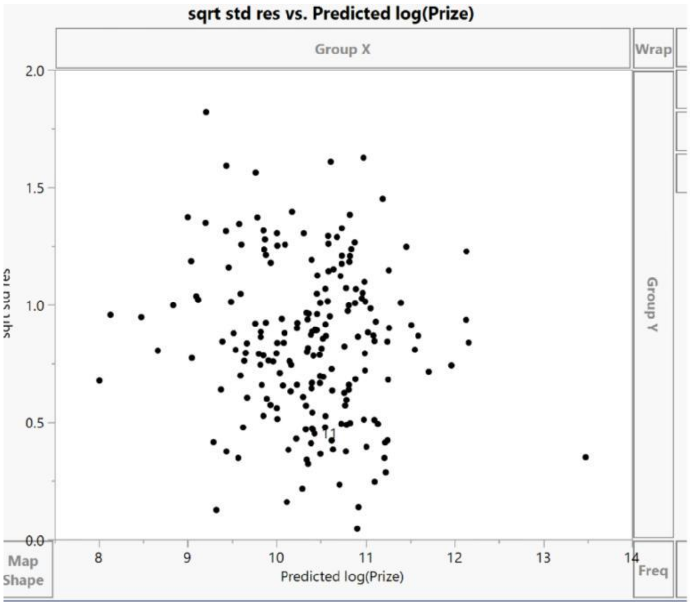
But, there is a linear relationship between PuttingAverage vs PuttsPerRound, and PuttingAverage vs Birdie Conversion% which should be investigated.

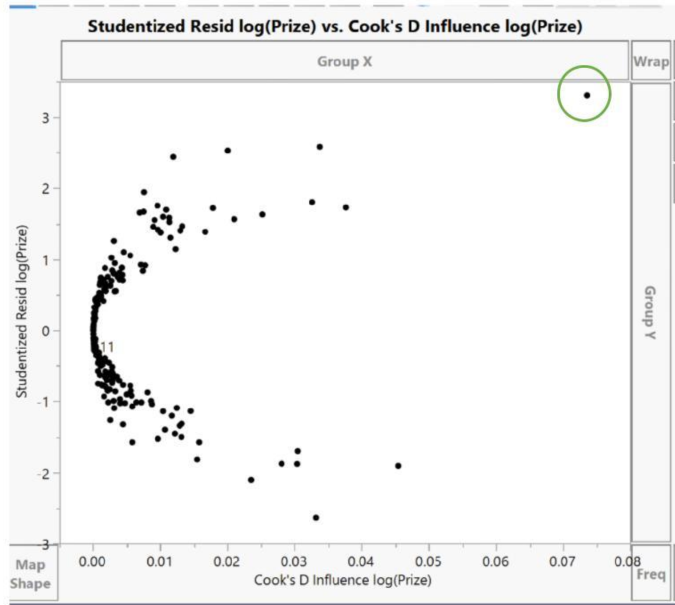

Scatterplot Matrix
DrivingAccuracy
GIR
PuttingAverage
BirdieConversion
SandSaves
Scrambling
PuttsPerRound
log(Prize)
1.86


Residual Normal Quantile Plot
log(Prize) Residual
Normal Quantile


Studentized Resid log(Prize) vs. Predicted log(Prize)
Group X
Wrap
Group Y
Studentized Resid log(Prize)
Predicted log(Prize)
Map Shape
Freq

Cook’s D plot looks almost normal except for the one in the top-right corner.





Part – C

The point 185 is investigated which is very influential due to high Cook’s D and has standardized residual value > 3. In addition to it, there are other points who have std res > 2 / <-2.

Part – D

VIF of the model is calculated. Some of the variables – PuttingAverage and PuttsPerRound have high VIF values which confirms the scatter plot matrix

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.1943003	7.777129	0.02	0.9801	.
DrivingAccuracy	-0.00353	0.011773	-0.30	0.7646	1.7966156
GIR	0.1993109	0.043817	4.55	<.0001*	6.2949685
PuttingAverage	-0.466304	6.905698	-0.07	0.9462	12.900789
BirdieConversion	0.1573409	0.040378	3.90	0.0001*	3.5118982
SandSaves	0.0151744	0.009862	1.54	0.1256	1.4615055
Scrambling	0.0515137	0.031788	1.62	0.1068	4.4702033
PuttsPerRound	-0.343131	0.473549	-0.72	0.4696	19.355667

Part- E

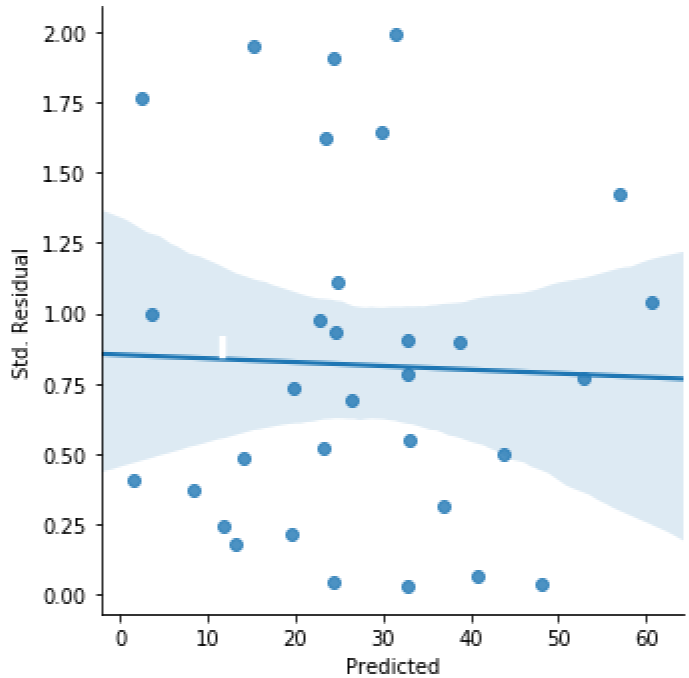
The model suffers from multi-colinearity. All the variables should not be removed at the same time. t-value and p-value cannot be taken as conclusive factors. The solution would be to remove one insignificant variable at a time. This can change the coefficient estimates, p-values of other insignificant variables. As a result, more variable can become significant.

Python solution of the same

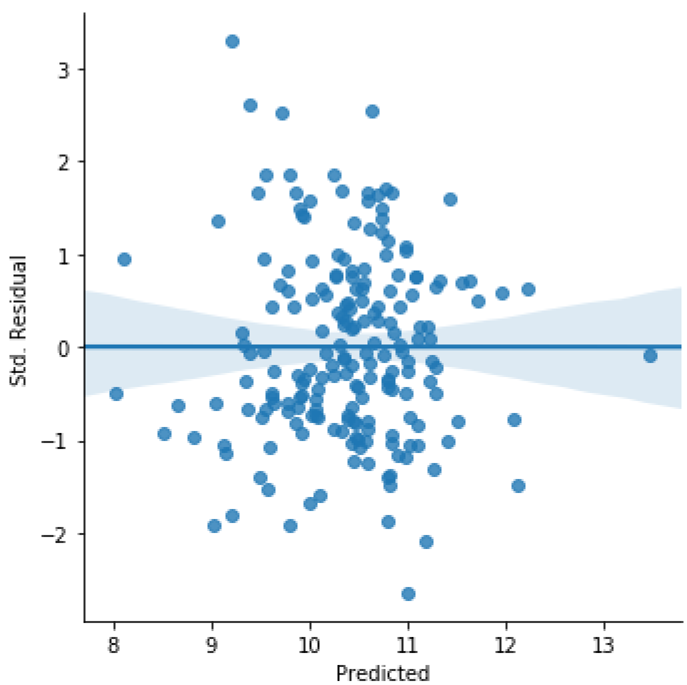
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             PrizeMoney   R-squared:                       0.407
Model:                            OLS   Adj. R-squared:                  0.381
Method:                 Least Squares   F-statistic:                     16.03
Date:                Fri, 08 Mar 2019   Prob (F-statistic):           6.31e-18
Time:                        22:25:28   Log-Likelihood:                -2395.2
No. Observations:                 196   AIC:                             4808.
Df Residuals:                     187   BIC:                             4838.
Df Model:                           8
Covariance Type:            nonrobust
======================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept        -1.148e+06    5.9e+05     -1.944      0.053   -2.31e+06    1.67e+04
DrivingAccuracy  -1845.4424    891.516     -2.070      0.040   -3604.163     -86.722
GIR               9301.2132   3450.304      2.696      0.008    2494.691    1.61e+04
PuttingAverage   -8.938e+04   5.34e+05     -0.167      0.867   -1.14e+06    9.64e+05
BirdieConversion  9957.9553   3284.341      3.032      0.003    3478.834    1.64e+04
SandSaves         1198.4418    747.689      1.603      0.111    -276.547    2673.430
Scrambling        4826.8211   2416.261      1.998      0.047      60.188    9593.454
BounceBack         616.8496   1583.833      0.389      0.697   -2507.626    3741.325
PuttsPerRound     7934.3442   3.65e+04      0.217      0.828   -6.41e+04    7.99e+04
==============================================================================
Omnibus:                      195.184   Durbin-Watson:                   1.942
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6293.342
Skew:                           3.687   Prob(JB):                         0.00
Kurtosis:                      29.763   Cond. No.                     2.25e+04
==============================================================================
```

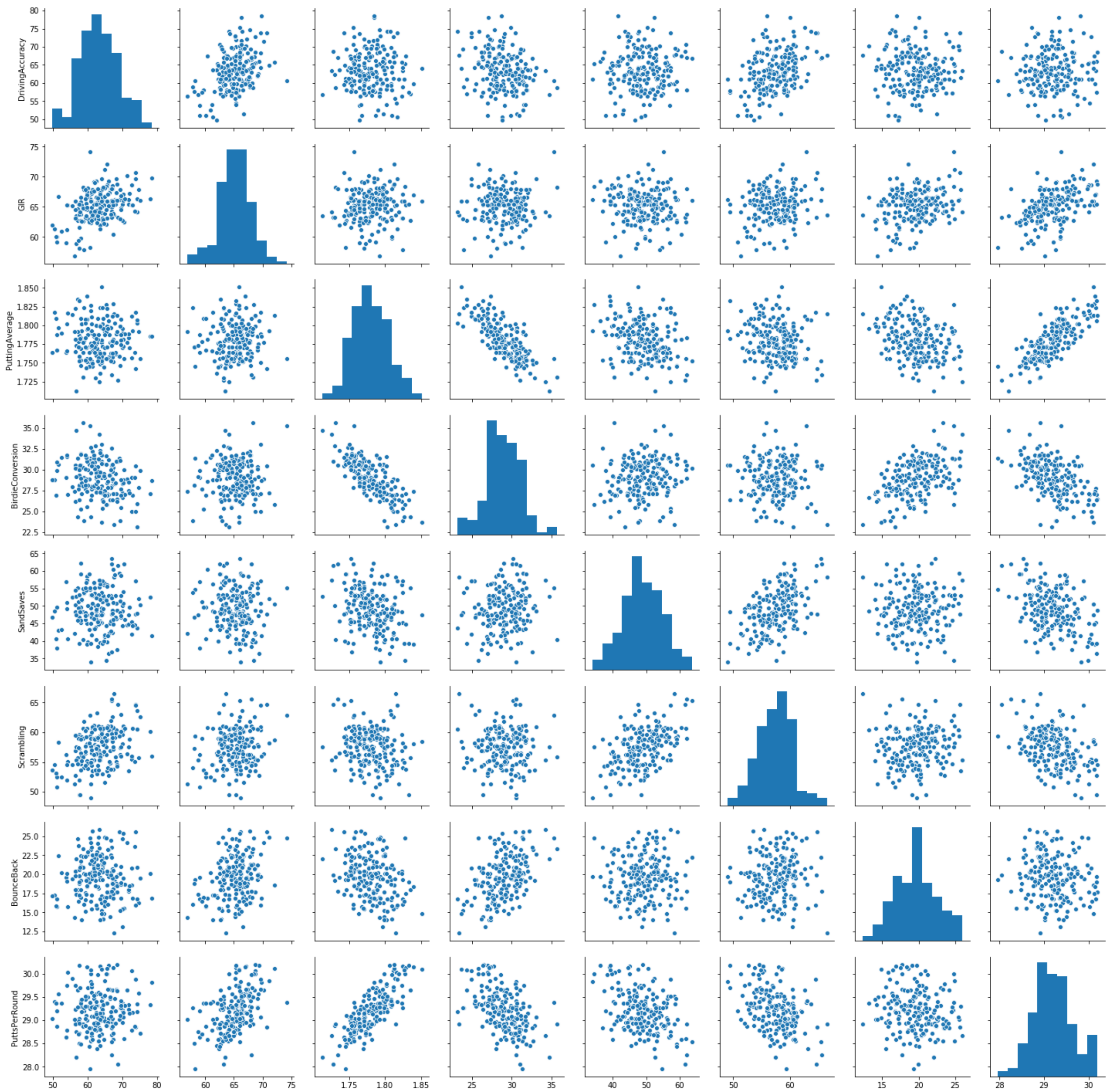
Summary After Transformation:

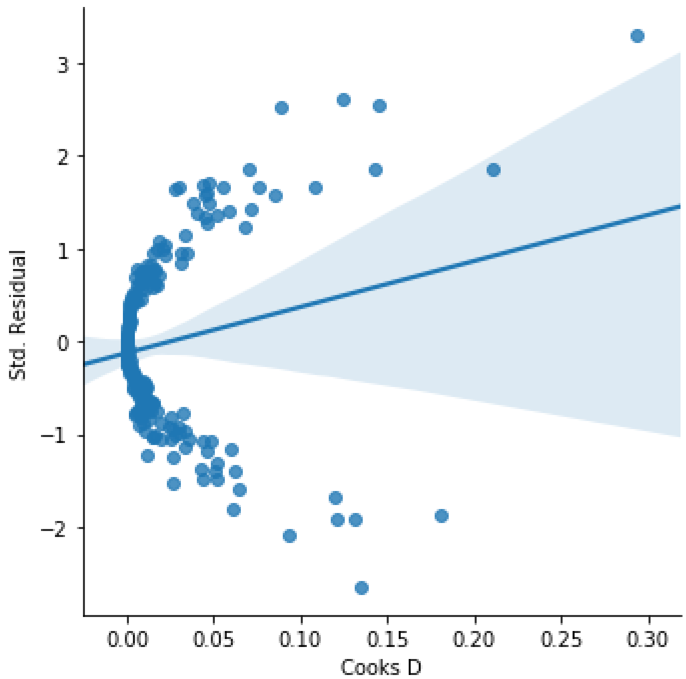
```
==============================================================================
Dep. Variable:             PrizeMoney   R-squared:                       0.559
Model:                            OLS   Adj. R-squared:                  0.540
Method:                 Least Squares   F-statistic:                     29.61
Date:                Fri, 08 Mar 2019   Prob (F-statistic):           1.53e-29
Time:                        22:36:38   Log-Likelihood:                -193.49
No. Observations:                 196   AIC:                             405.0
Df Residuals:                     187   BIC:                             434.5
Df Model:                           8
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept            0.6007      7.810      0.077      0.939     -14.806      16.008
DrivingAccuracy     -0.0038      0.012     -0.318      0.750      -0.027       0.020
GIR                  0.1906      0.046      4.176      0.000       0.101       0.281
PuttingAverage      -1.4532      7.061     -0.206      0.837     -15.382      12.475
BirdieConversion     0.1463      0.043      3.368      0.001       0.061       0.232
SandSaves            0.0156      0.010      1.573      0.117      -0.004       0.035
Scrambling           0.0535      0.032      1.675      0.096      -0.010       0.117
BounceBack           0.0145      0.021      0.693      0.489      -0.027       0.056
PuttsPerRound       -0.2804      0.483     -0.581      0.562      -1.233       0.672
==============================================================================
Omnibus:                        4.691   Durbin-Watson:                   1.814
Prob(Omnibus):                  0.096   Jarque-Bera (JB):                4.478
Skew:                           0.369   Prob(JB):                        0.107
Kurtosis:                       3.070   Cond. No.                     2.25e+04
==============================================================================
```



Variance of error terms now looks random.




DrivingAccuracy
GIR
PuttingAverage
BirdieConversion
SandSaves
Scrambling
BounceBack
PuttsPerRound



3
2
1
0
−1
−2
Std. Residual
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Cooks D