DSC 423: Data Analytics and Regression

Assignment 08

Name: Adarsh Shankar

Student Id: 2117611

Honor Statement: "I have completed this work independently. The solutions given are entirely my own work."

1. Short Essay (10 points). Read the short PDF on George Box. Explain in your own words the significance of "all models are wrong, but some are useful" as if you were interviewing for job in data science.

2. Previously, you used the PGA tour dataset to predict Prize Money. Use a log transformation to transform Prize Money into a new response variable. Apply your knowledge of regression analysis to fit a regression model using the remaining predictors in your dataset. 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.

a. **(10 points) Check for multicollinear. Explain your process. In our final model, multicollinearity exists. We can tell that GIR, PABB, PuttingAverage, BB2, SBBB, PuttsPerRound, ADDPA, G2, ADDPPR, PA2 are all greater than 10.**

```
> model3 <- lm(log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound + G2 + PA2 + BC2 + BB2 + ADDPA + ADDPPR + DASB +
 PABB + SBBB, data = d)
> summary(model3)

Call:
lm(formula = log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound +
    G2 + PA2 + BC2 + BB2 + ADDPA + ADDPPR + DASB + PABB + SBBB,
    data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.44164 -0.48010 -0.09378  0.37145  1.91996

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.366e+02  1.863e+02  -2.344 0.020171 *
GIR             1.675e+00  5.261e-01   3.184 0.001705 **
PuttingAverage  6.560e+02  2.261e+02   2.902 0.004166 **
PuttsPerRound  -1.338e+01  5.420e+00  -2.468 0.014502 *
G2             -1.132e-02  4.067e-03  -2.783 0.005947 **
PA2            -1.209e+02  5.766e+01  -2.097 0.037370 *
BC2             2.991e-03  8.274e-04   3.615 0.000388 ***
BB2             1.085e-02  4.559e-03   2.380 0.018354 *
ADDPA          -7.591e-01  3.115e-01  -2.437 0.015776 *
ADDPPR          4.573e-02  1.901e-02   2.405 0.017173 *
DASB           -5.354e-04  2.390e-04  -2.240 0.026314 *
PABB           -3.980e-01  1.119e-01  -3.557 0.000478 ***
SBBB            5.089e-03  1.630e-03   3.122 0.002089 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6354 on 183 degrees of freedom
Multiple R-squared:  0.6057,    Adjusted R-squared:  0.5798
F-statistic: 23.42 on 12 and 183 DF,  p-value: < 2.2e-16


> vif(model3)
          GIR PuttingAverage  PuttsPerRound            G2           PA2           BC2           BB2
   990.930674   15095.465044    2768.725043   997.366537  12471.612566      5.417773    123.283788
        ADDPA         ADDPPR           DASB         PABB         SBBB
 14669.159718   17146.356517       5.120463   143.318351     41.241249
```

d$G2 <- d$GIR^2 d$PA2 <- d$PuttingAverage^2 d$BC2 <- d$BirdieConversion^2 d$PABB <- d$PuttingAverage * d$BounceBack d$BB2 <- d$BounceBack^2 d$SBBB <- d$Scrambling * d$BounceBack d$ADDPA <- d$AveDrivingDistance * d$PuttingAverage d$ADDPPR <- d$AveDrivingDistance * d$PuttsPerRound

b. (10 points) Compare this model to the one you made in the previous assignment. How did performing a log transformation impact the quality of the model? Why?

Last Assignment Final Model:

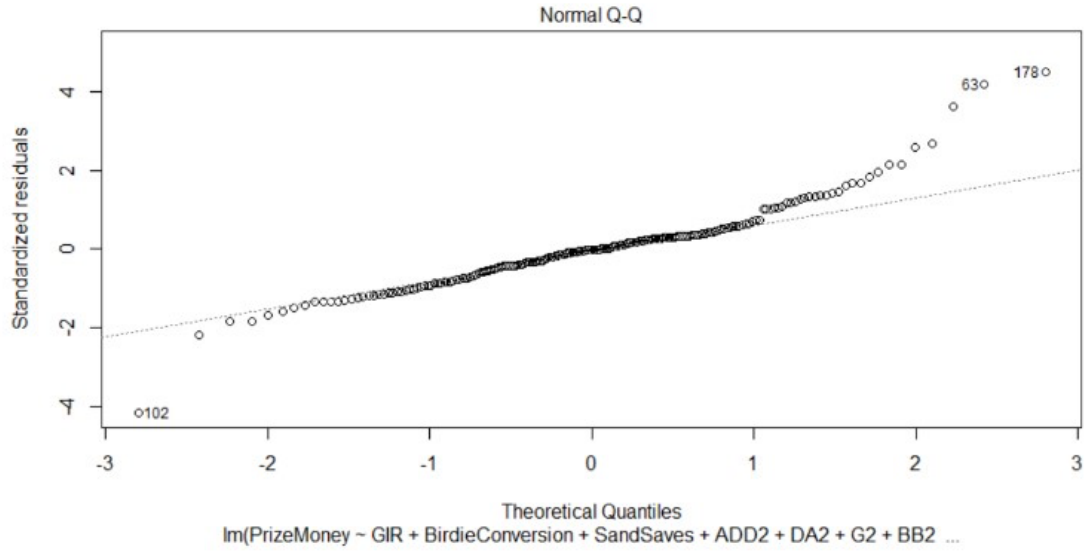
lafm <- lm(PrizeMoney ~ GIR + BirdieConversion + SandSaves + ADD2 + DA2 + G2 + BB2 +

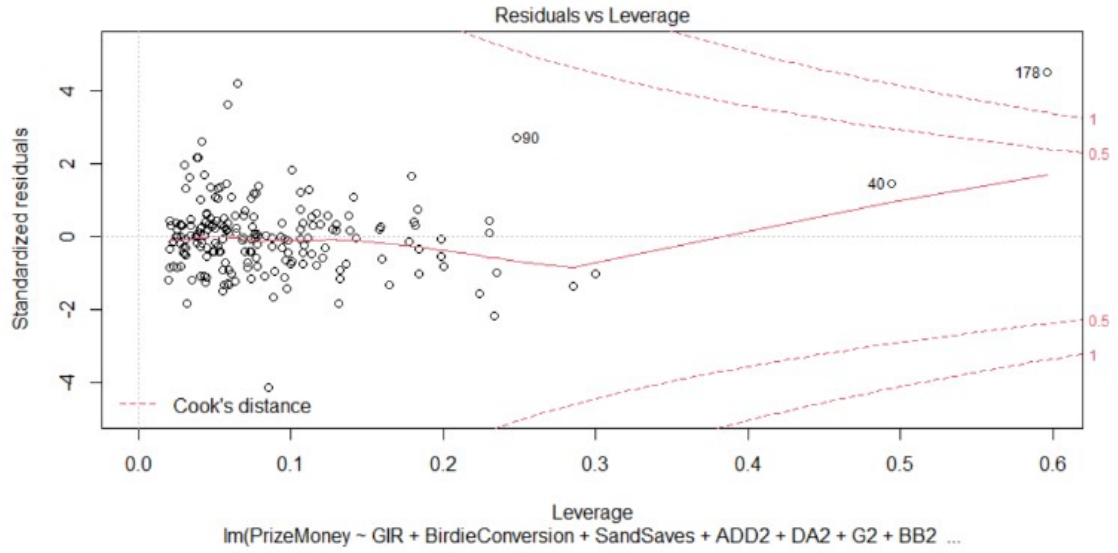
ADDBC + ADDSS + DAG + GPA + GBC + PASB + PABB + SSSB + SBBB, data = d)

summary(lafm)

```
> lafm <- lm(PrizeMoney ~ GIR + BirdieConversion + SandSaves + ADD2 + DA2 + G2 + BB2 + ADDBC + ADDSS + DAG + GPA + GBC
 + PASB + PABB + SSSB + SBBB, data = d)
> summary(lafm)

Call:
lm(formula = PrizeMoney ~ GIR + BirdieConversion + SandSaves +
    ADD2 + DA2 + G2 + BB2 + ADDBC + ADDSS + DAG + GPA + GBC +
    PASB + PABB + SSSB + SBBB, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-142234  -19973    -990   12435  145195

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.245e+07  1.210e+06  10.291  < 2e-16 ***
GIR           -2.026e+05  2.966e+04  -6.830 1.28e-10 ***
BirdieConversion -2.481e+05  4.585e+04  -5.413 1.97e-07 ***
SandSaves     -1.164e+05  2.315e+04  -5.027 1.20e-06 ***
ADD2          -4.187e+01  9.145e+00  -4.579 8.72e-06 ***
DA2            2.674e+02  8.536e+01   3.133 0.002024 **
G2             8.699e+02  2.497e+02   3.484 0.000620 ***
BB2            7.458e+02  2.587e+02   2.883 0.004422 **
ADDBC          3.308e+02  1.654e+02   1.999 0.047088 *
ADDSS          2.945e+02  6.291e+01   4.681 5.62e-06 ***
DAG           -5.519e+02  1.694e+02  -3.259 0.001339 **
GPA            3.466e+04  7.522e+03   4.608 7.69e-06 ***
GBC            2.502e+03  4.154e+02   6.023 9.52e-09 ***
PASB          -2.255e+04  4.749e+03  -4.748 4.19e-06 ***
PABB          -4.373e+04  1.116e+04  -3.918 0.000127 ***
SSSB           5.603e+02  1.336e+02   4.192 4.33e-05 ***
SBBB           8.536e+02  3.048e+02   2.800 0.005669 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35790 on 179 degrees of freedom
Multiple R-squared:  0.7121,    Adjusted R-squared:  0.6864
F-statistic: 27.68 on 16 and 179 DF,  p-value: < 2.2e-16
```

d$ADD2 <- d$AveDrivingDistance^2 d$DA2 <- d$DrivingAccuracy^2 d$G2 <- d$GIR^2 d$BB2 <- d$BounceBack^2 d$ADDBC <- d$AveDrivingDistance * d$BirdieConversion d$ADDSS <- d$AveDrivingDistance * d$SandSaves d$DAG <- d$DrivingAccuracy * d$GIR d$GPA <- d$GIR * d$PuttingAverage d$GBC <- d$GIR * d$BirdieConversion d$PASB <- d$PuttingAverage * d$Scrambling d$PABB <- d$PuttingAverage * d$BounceBack d$SSSB <- d$SandSaves * d$Scrambling d$SBBB <- d$Scrambling * d$BounceBack

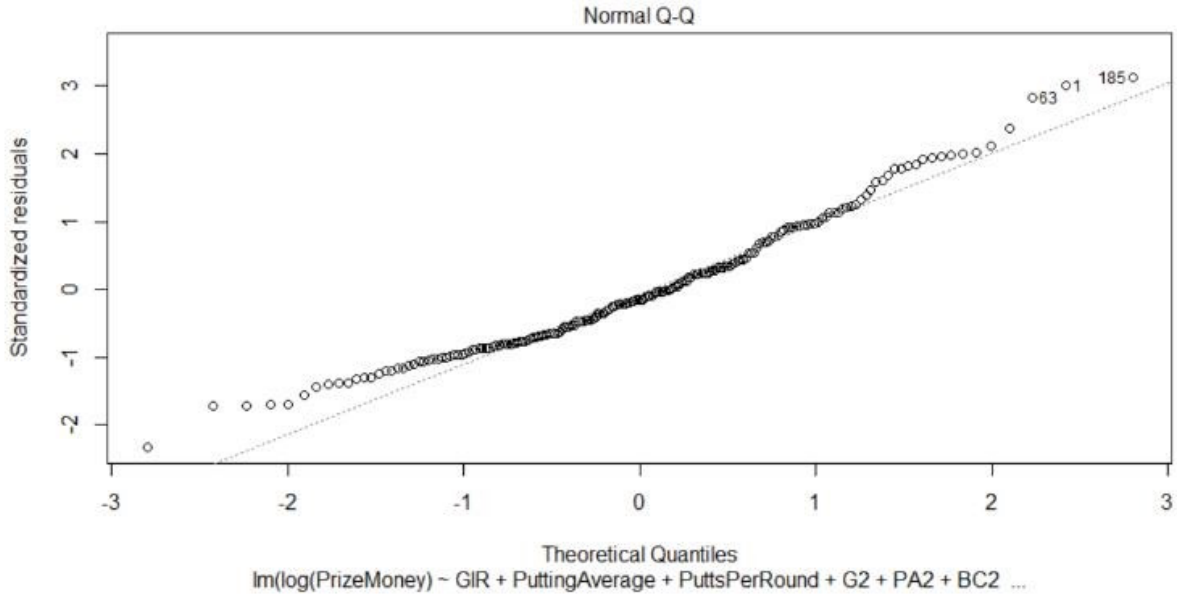


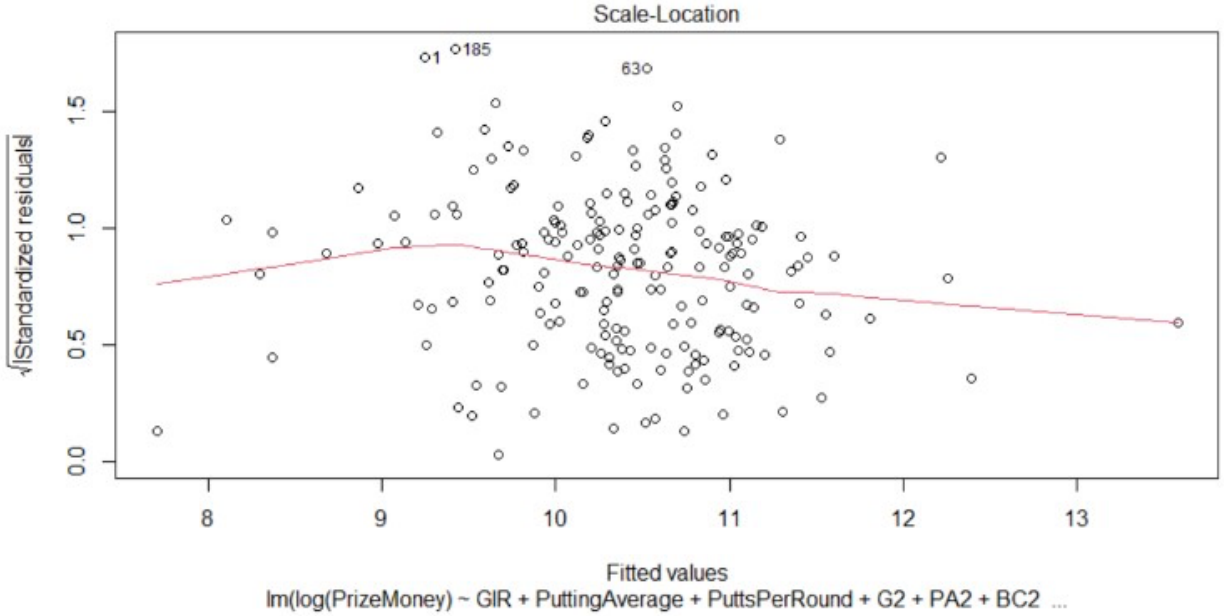


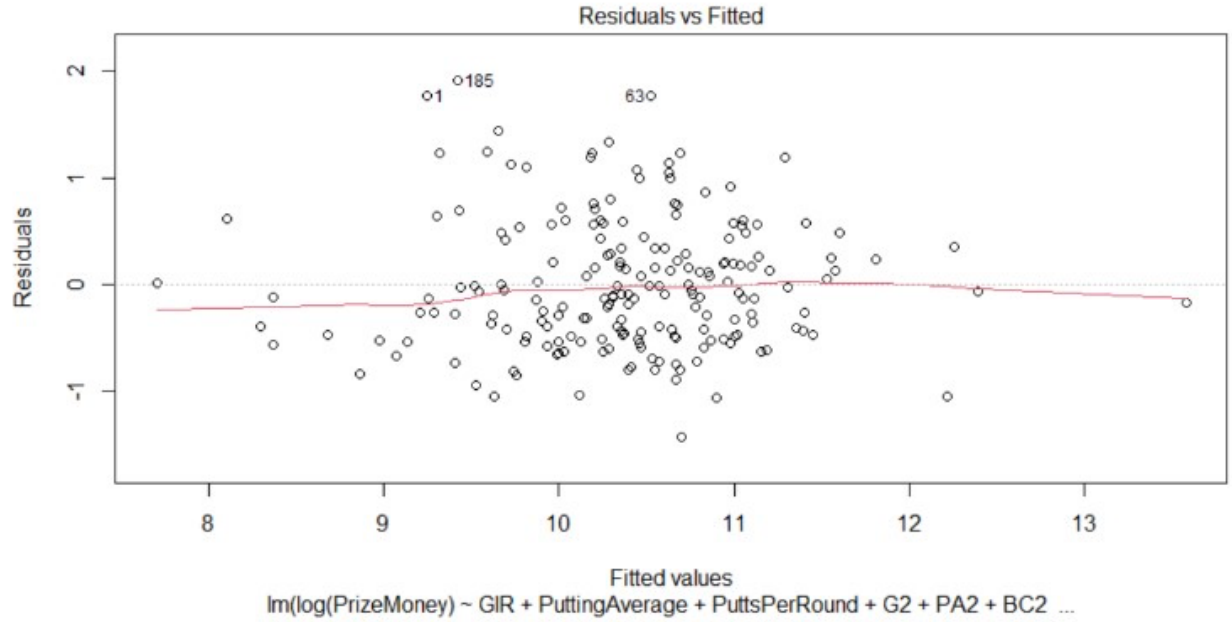
Present assignment Model:

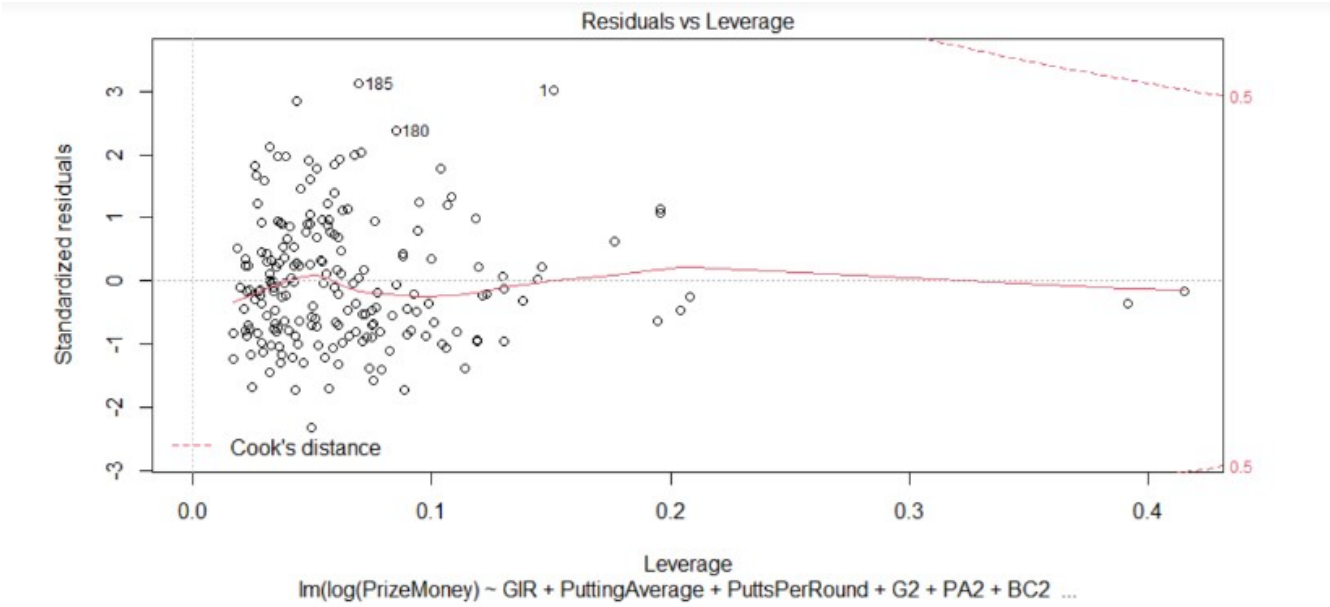
```
> model3 <- lm(log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound + G2 + PA2 + BC2 + BB2 + ADDPA + ADDPPR + DASB +
 PABB + SBBB, data = d)
> summary(model3)

Call:
lm(formula = log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound +
    G2 + PA2 + BC2 + BB2 + ADDPA + ADDPPR + DASB + PABB + SBBB,
    data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.44164 -0.48010 -0.09378  0.37145  1.91996

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.366e+02  1.863e+02  -2.344 0.020171 *
GIR             1.675e+00  5.261e-01   3.184 0.001705 **
PuttingAverage  6.560e+02  2.261e+02   2.902 0.004166 **
PuttsPerRound  -1.338e+01  5.420e+00  -2.468 0.014502 *
G2             -1.132e-02  4.067e-03  -2.783 0.005947 **
PA2            -1.209e+02  5.766e+01  -2.097 0.037370 *
BC2             2.991e-03  8.274e-04   3.615 0.000388 ***
BB2             1.085e-02  4.559e-03   2.380 0.018354 *
ADDPA          -7.591e-01  3.115e-01  -2.437 0.015776 *
ADDPPR          4.573e-02  1.901e-02   2.405 0.017173 *
DASB           -5.354e-04  2.390e-04  -2.240 0.026314 *
PABB           -3.980e-01  1.119e-01  -3.557 0.000478 ***
SBBB            5.089e-03  1.630e-03   3.122 0.002089 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6354 on 183 degrees of freedom
Multiple R-squared:  0.6057,    Adjusted R-squared:  0.5798
F-statistic: 23.42 on 12 and 183 DF,  p-value: < 2.2e-16
```

We can use summary(model) and plot to compare the two models (model). As can be observed from the comparison, the adj-R2 value of the prior model without the log transformation is higher. On the other hand, the present model log transforms the normalQQ data. The graph looks like a direct diagonal In this case, it shows a straight line, indicating that the model's distribution is linearly normal. As a result, there are also less linked outliers. The final model's log function transformation on PrizeMoney's (y value) forces us to select alternative x variables at the same time.




Scale-Location
√|Standardized residuals|
1
185
63
0.0
0.5
1.0
1.5
8
9
10
11
12
13
Fitted values
lm(log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound + G2 + PA2 + BC2 ...

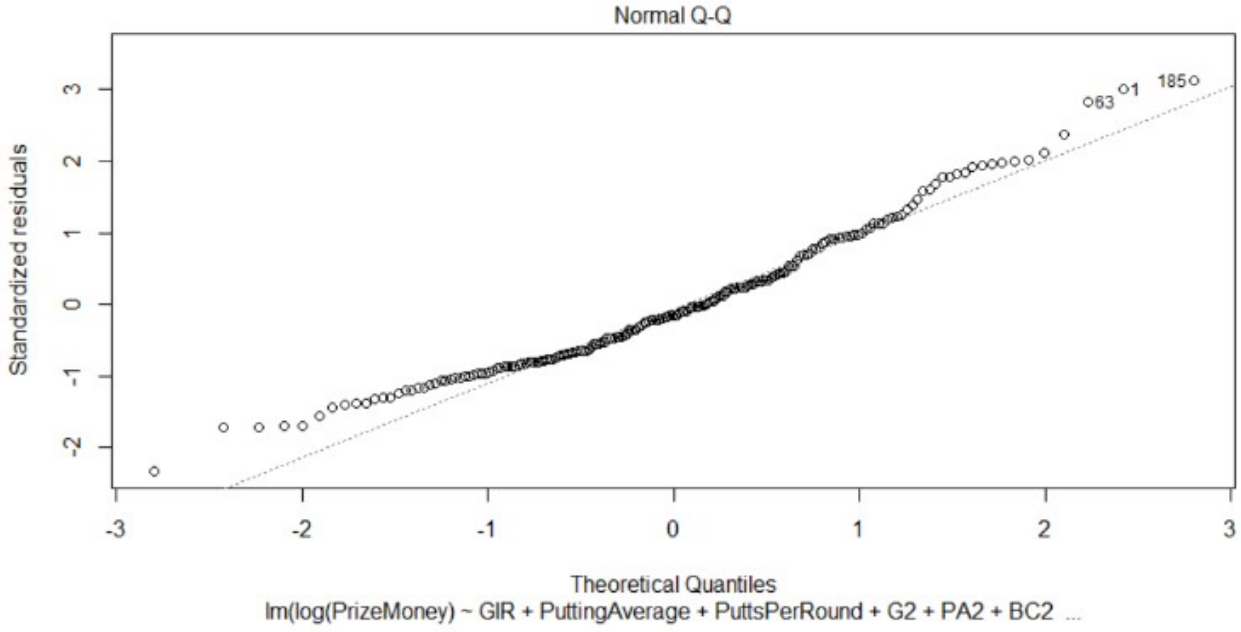

Residuals vs Fitted
Residuals
1
185
63
-1
0
1
2
8
9
10
11
12
13
Fitted values
lm(log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound + G2 + PA2 + BC2 ...

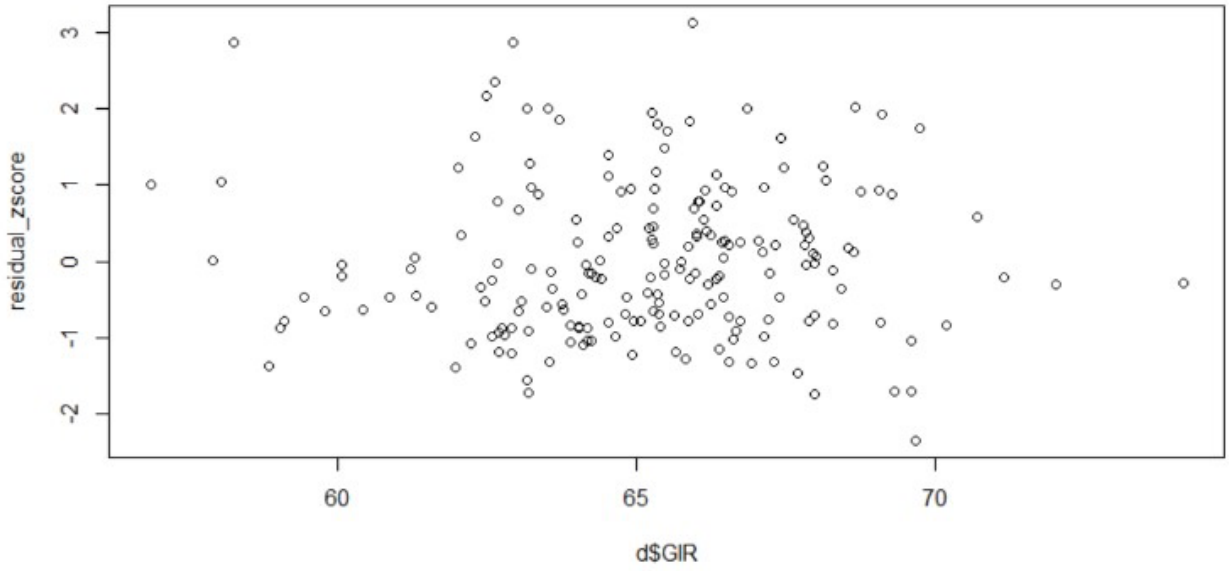


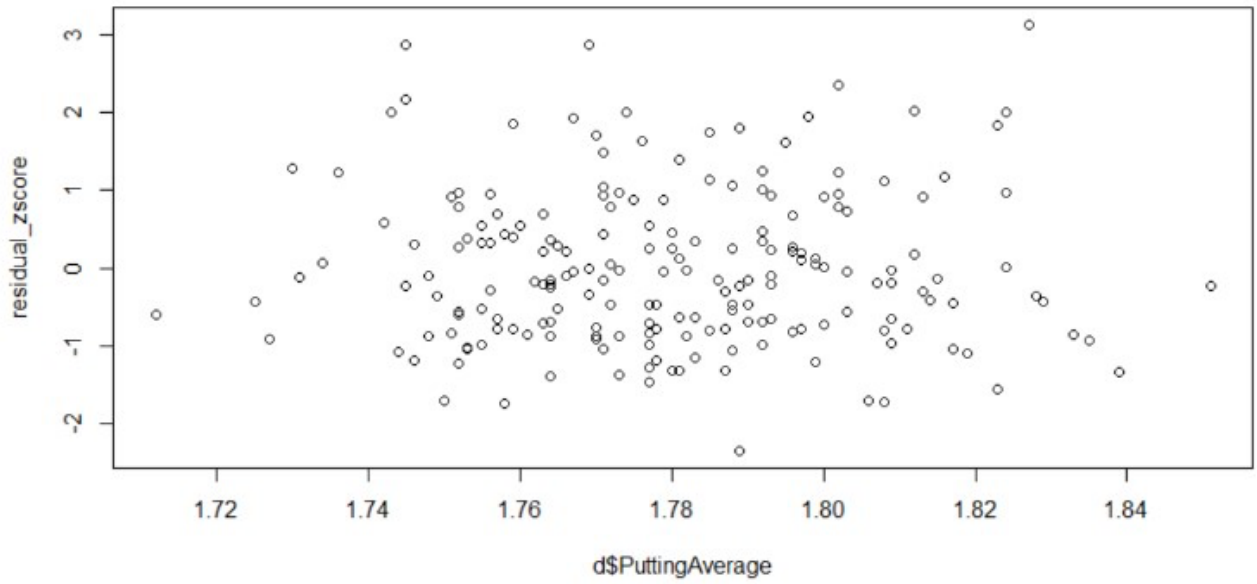
c. (10 points) Analyze and discuss the residual plots.

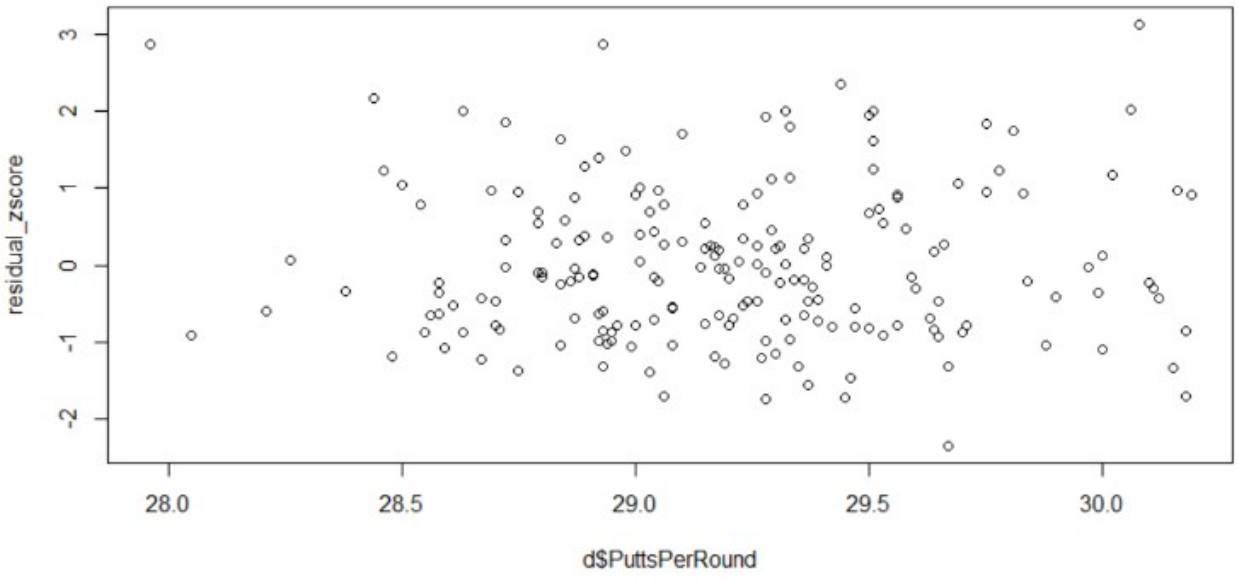
In our current log model of studentized residual graph, the model is normally distributed along a linear line. The zero line has x variable graphs scattered about it at random. They stand for independence and continual variance. There might be outliers in the graphs of the x variables because the studentized residuals are either more than 3 or smaller than -3.

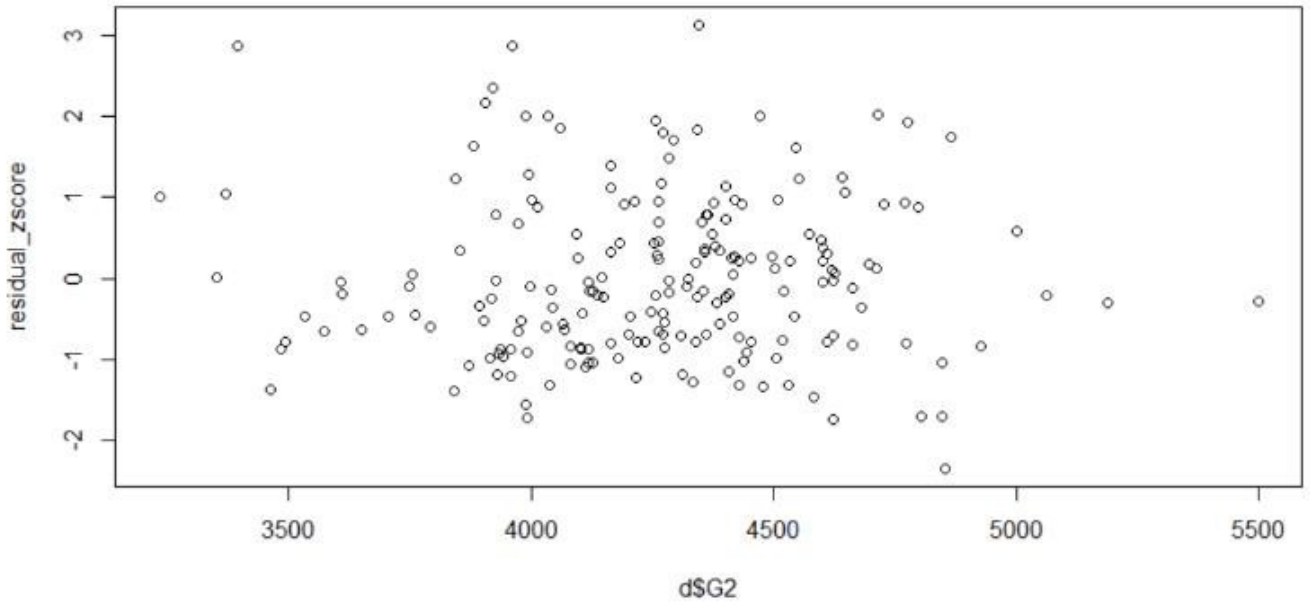
```
> model3$residuals
          1           2           3           4           5           6           7           8
 1.7639270581  1.1866988553 -0.4802417804 -0.8989371504 -0.5227592650  0.5416687437  1.2431089334  0.7553950789
          9          10          11          12          13          14          15          16
 1.1910539138 -0.7291568037 -0.0978786571  0.1498361711  0.1286138761 -0.2843161563  0.1515819327 -0.4907090553
         17          18          19          20          21          22          23          24
-0.6408560886  0.6437562299 -0.0936903338 -0.4800548183 -0.0283255435  0.1290567022  0.6903514043 -0.4290391155
         25          26          27          28          29          30          31          32
-0.5579979016 -0.5118318331 -0.4368928555 -0.4813333204  0.2784346116  0.9981166664 -0.2530784110  0.9933067729
         33          34          35          36          37          38          39          40
 1.1307034404  0.0670651778 -0.6260177741  0.5751899290  0.0251946084 -0.3234961085 -1.0456420590 -0.0816345422
         41          42          43          44          45          46          47          48
-0.0100207364 -0.5422049729  0.1491601179  0.2052089952 -0.3657446871 -0.1540184924 -1.4416404584 -0.0228696299
         49          50          51          52          53          54          55          56
 0.7921705071 -0.7824101285  0.2402238106  0.5639914402  0.1768248640 -0.3939688638 -0.3956104873 -0.3401119644
         57          58          59          60          61          62          63          64
-0.0624706474  1.1412095618 -0.4403715802  0.7191728396  0.6979922058  0.3316460024  1.7637332678 -0.5399228818
         65          66          67          68          69          70          71          72
-0.5438394116  0.1069304675 -0.5353639693  0.2869146452  0.0005610607  0.0100668708 -0.2907808298 -0.0659520147
         73          74          75          76          77          78          79          80
 0.4801810232  0.5877711763  0.6472967752  0.4852800576 -0.6355835670 -0.2742704902  0.1144040017  0.4184794069
         81          82          83          84          85          86          87          88
-1.0607450440 -0.8131471663 -0.1829222686  1.1015818111 -0.4000466816 -0.1285425340  0.0757851607 -0.8050391193
         89          90          91          92          93          94          95          96
-0.1419281473  0.3557800320 -0.8608012356  1.0723455065 -0.4865481983  0.0685488951 -0.1399333488 -0.2859507770
         97          98          99         100         101         102         103         104
-0.3997182128 -0.6090423043  0.3372825610 -0.1309170841 -0.5734362476 -1.0513660495 -0.2167600970  1.0506000505
        105         106         107         108         109         110         111         112
 0.5345134764 -0.4793482377  0.7596886318  0.5655057414 -0.1094391647 -0.5024287119  0.4258320169 -0.7045296273
        113         114         115         116         117         118         119         120
-0.0328713754 -0.3228159923 -0.8464240417  0.1313877407  0.4812119306 -0.7529244044 -0.7446044580  0.2077954978
        121         122         123         124         125         126         127         128
-0.2898503951 -1.0734014349  0.5956303576 -0.4515993583 -0.8241411423 -0.7253846662 -0.6751460746 -0.9574847933
        129         130         131         132         133         134         135         136

 0.1699151525 -0.4077902913 -0.2635641681 -0.3398294337 -0.1313868424 -0.1204342016  0.2625955437  0.2646898410
        137         138         139         140         141         142         143         144
-0.6551274800 -0.2146511947  0.1437597485 -0.4860353726  1.2258175000 -0.0745927242  0.9083104579 -0.1845150699
        145         146         147         148         149         150         151         152
 0.2158022027  0.1906348622 -0.8114973303 -0.2891181836 -0.6026919748  1.3295095676 -0.5301567228 -0.4207595158
        153         154         155         156         157         158         159         160
 0.8575441168 -0.1185083076  0.5993104244 -0.0979715969 -0.4264842986 -0.5602416386  0.1985957398  0.5648911625
        161         162         163         164         165         166         167         168
 0.7486027412 -0.5199794545  0.3374152169 -0.2199735544 -0.4651688857  0.5957064491 -0.6022881913 -0.6355650919
        169         170         171         172         173         174         175         176
-0.0938761092 -0.5899878186 -0.0176061003  0.0437191334  0.5794048830  1.2320396341 -0.0626097286  0.1649485873
        177         178         179         180         181         182         183         184
-0.1423661810 -0.1739751156  0.4305635705  1.4410923387  0.0270057194 -0.4925411728 -0.6592912670  0.6149638944
        185         186         187         188         189         190         191         192
 1.9199628541  0.1967304647 -0.3478287047  0.5774157075  0.4445769641 -0.3676144952 -0.2630068714  0.2349784153
        193         194         195         196
-0.5419898174 -0.0121673134 -0.0214814002  1.2286858935
> sum(model3$residuals)
[1] -1.720846e-15
> mean = mean(model3$residuals)
> sd = sd(model3$residuals)
> residual_zscore = (model3$residuals - mean)/sd
> durbinWatsonTest(model3)
 lag Autocorrelation D-W Statistic p-value
   1      0.07924057      1.778971    0.12
 Alternative hypothesis: rho != 0
> plot(d$GIR, residual_zscore)
> plot(d$PuttingAverage, residual_zscore)
> plot(d$PuttsPerRound, residual_zscore)
> plot(d$G2, residual_zscore)
> plot(d$PA2, residual_zscore)
> plot(d$BC2, residual_zscore)
> plot(d$BB2, residual_zscore)
> plot(d$ADDPA, residual_zscore)
> plot(d$ADDPPR, residual_zscore)
> plot(d$DASB, residual_zscore)
> plot(d$PABB, residual_zscore)
> plot(d$SBBB, residual_zscore)
> plot(model3)
```

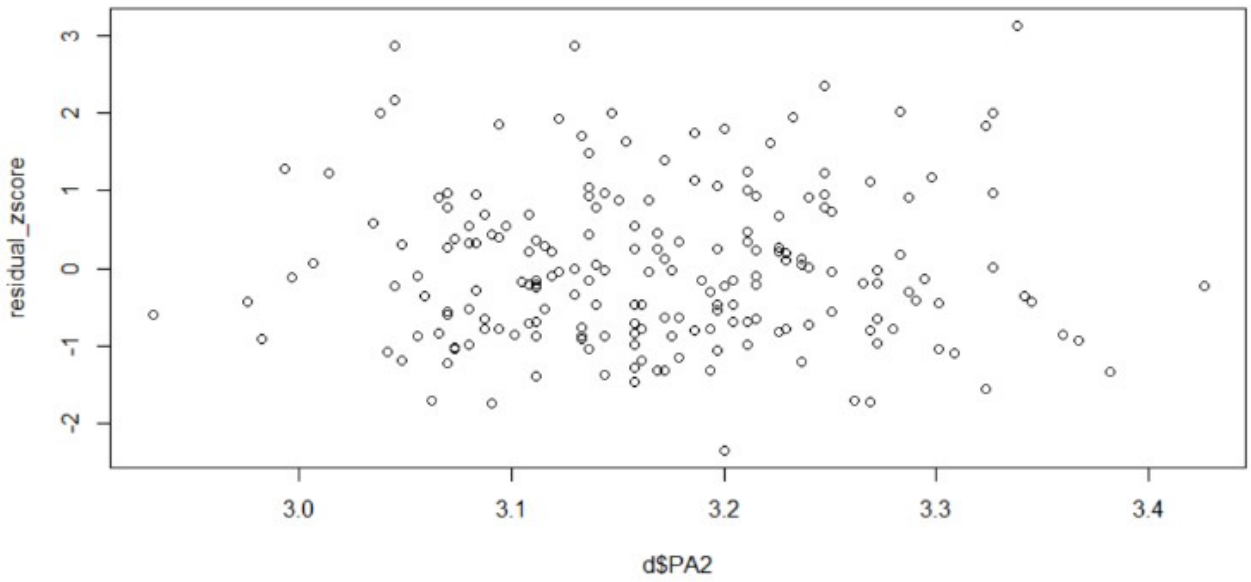

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound + G2 + PA2 + BC2 ...
63
1
185

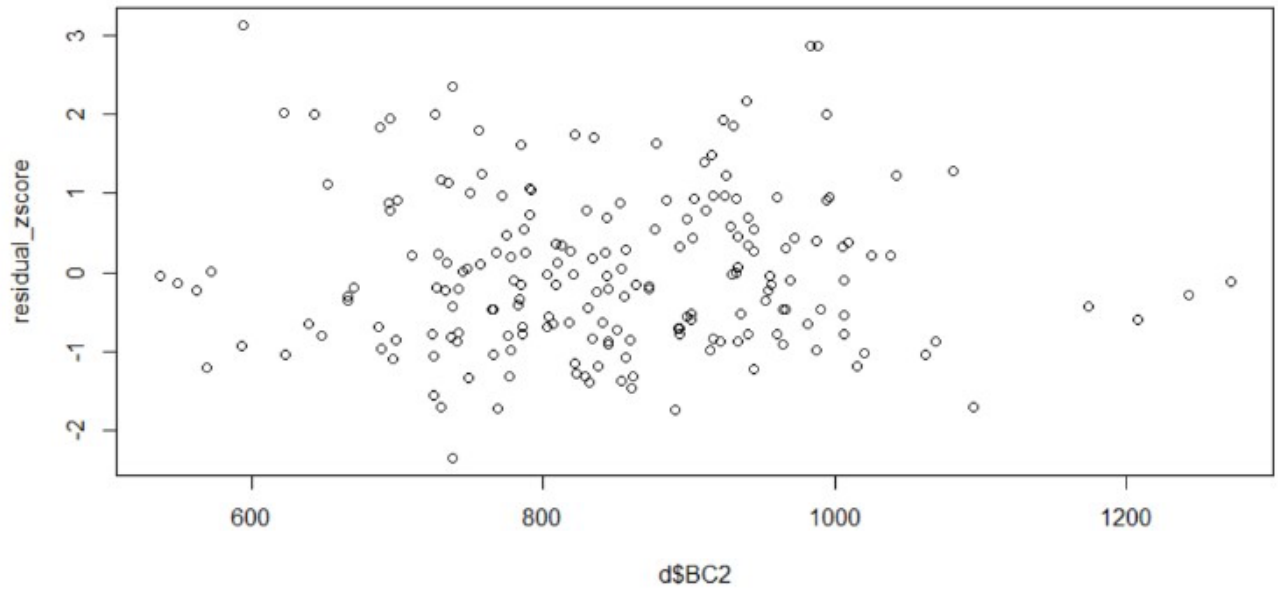

residual_zscore
d$GIR

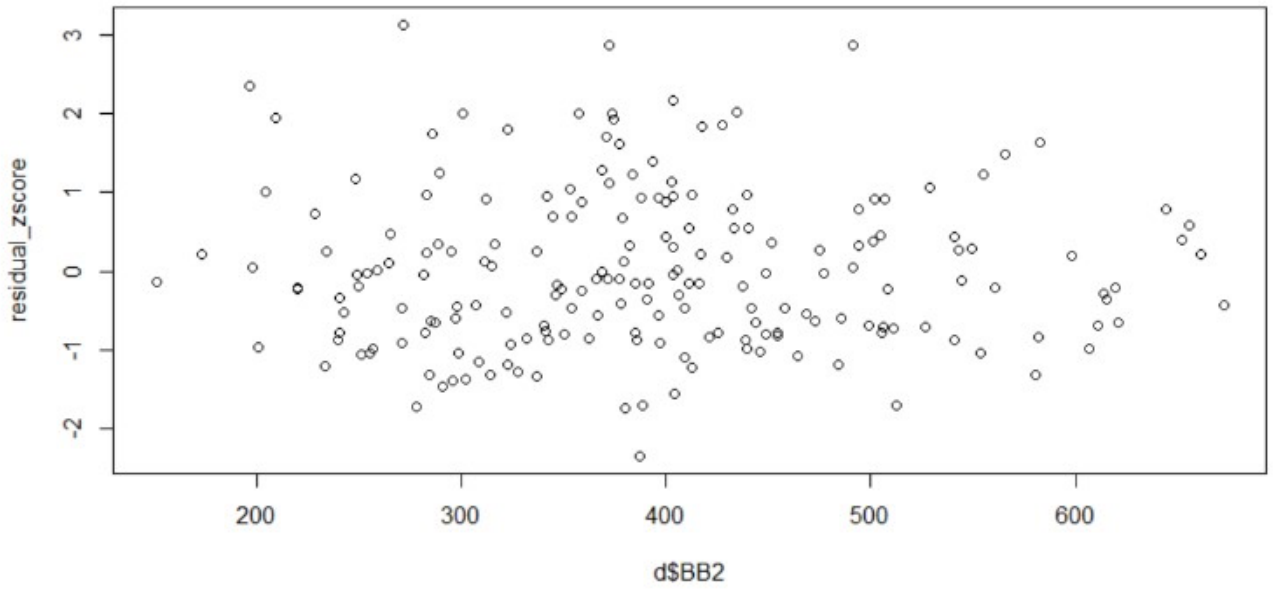

residual_zscore
3
2
1
0
-1
-2
1.72
1.74
1.76
1.78
1.80
1.82
1.84
d$PuttingAverage

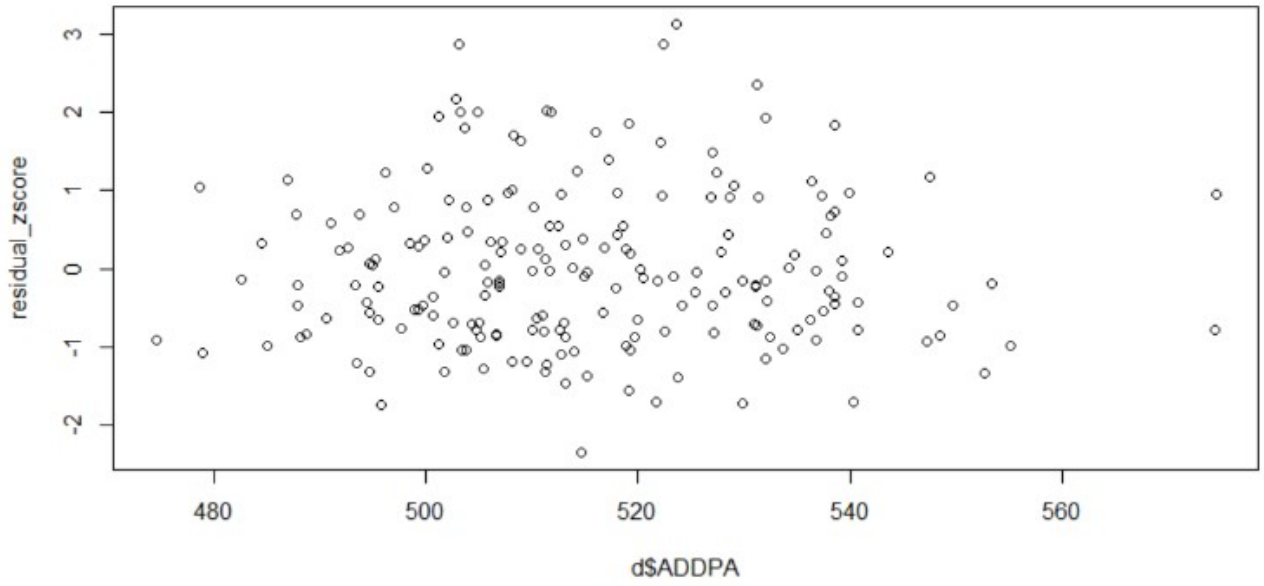

residual_zscore
3
2
1
0
-1
-2
28.0
28.5
29.0
29.5
30.0
d$PuttsPerRound

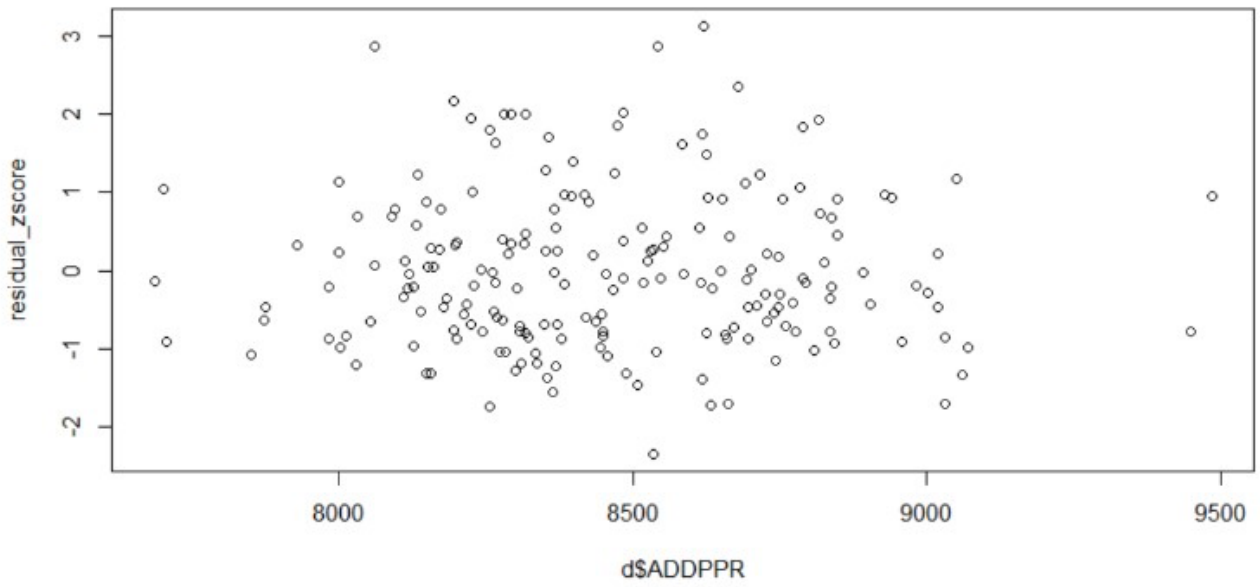

residual_zscore
3
2
1
0
-1
-2
3500
4000
4500
5000
5500
d$G2

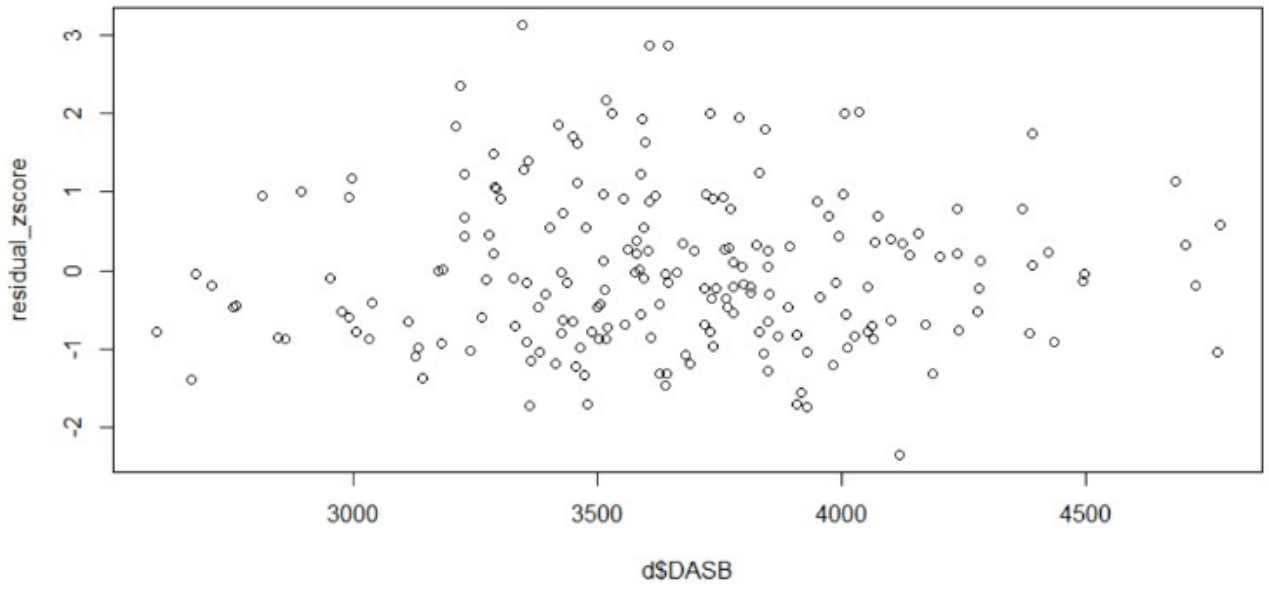

residual_zscore
3
2
1
0
-1
-2
3.0
3.1
3.2
3.3
3.4
d$PA2

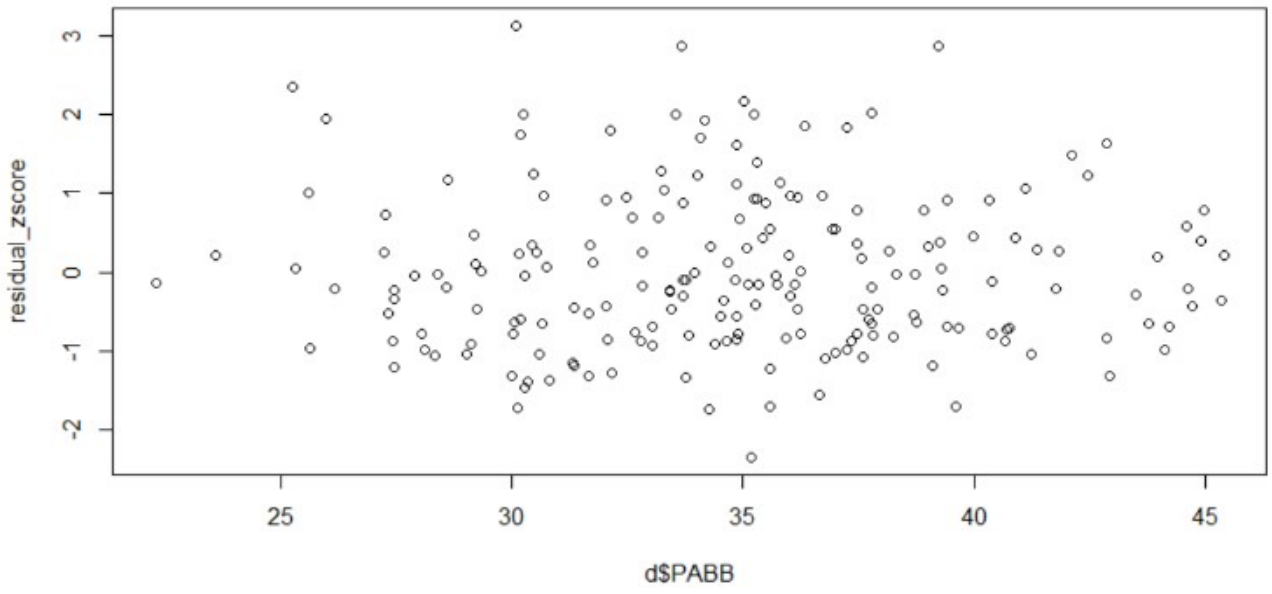

residual_zscore
3
2
1
0
-1
-2
600
800
1000
1200
d$BC2

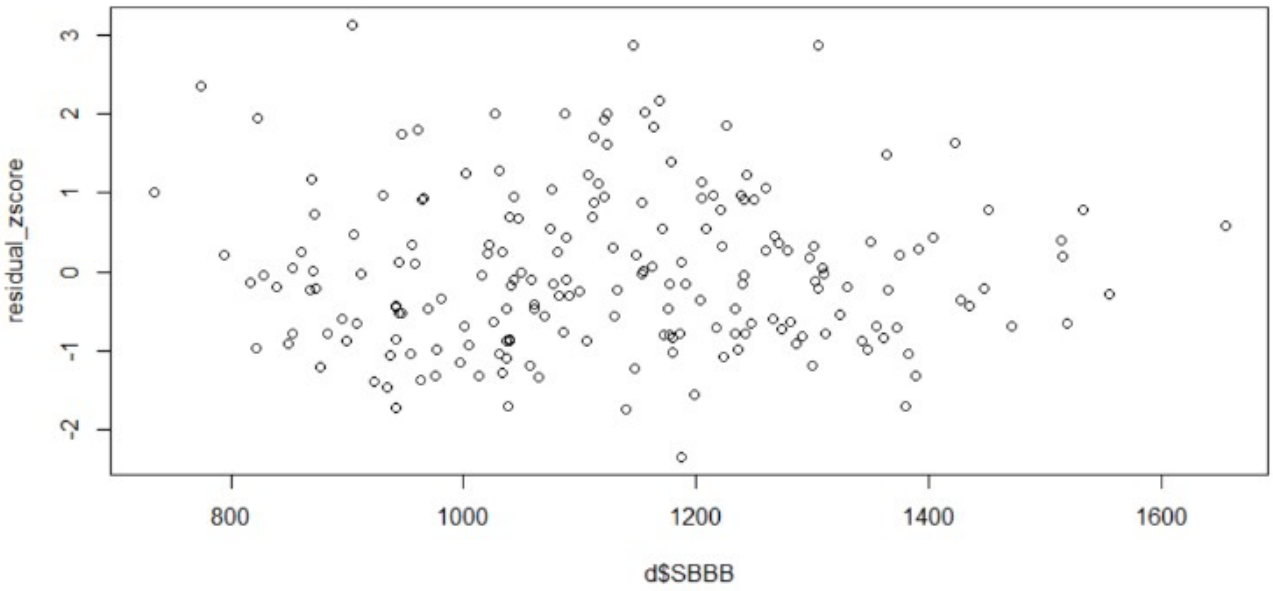

residual_zscore
3
2
1
0
-1
-2
200
300
400
500
600
d$BB2


residual_zscore
3
2
1
0
-1
-2
480
500
520
540
560
d$ADDPA


residual_zscore
3
2
1
0
-1
-2
8000
8500
9000
9500
d$ADDPPR


residual_zscore
-2
-1
0
1
2
3
3000
3500
4000
4500
d$DASB


residual_zscore
-2
-1
0
1
2
3
25
30
35
40
45
d$PABB



d. (10 points) 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. Discuss your answer.

Outliers are data points that deviate from the overall pattern by a large margin. When compared to other values, it can have extreme X or Y values, or both. An outlier that affects the slope of the regression line is known as an influential point. Calculate the regression equation with and without the outlier to see if the outlier has an impact. In our figures, there are likely outliers in the +3 range. We can deal with this by removing the outlier observations and running the model again. Examine the predictors' adj-R2, residual plots, and p-values to check to see if they've improved. Remove the influential point that was highlighted by nearly all indicators. Examine the predictors' adj-R2, residual plots, and pvalues. Check to see if they've improved. If it does, include it in your observations. doesn't. Rerun until adj-R2, the goodness of fit test, the residuals, and the p-values for each predictor are all in order. The overall goodness of fit test shows that at least one predictor is strongly correlated with Y if AdjR2 increases, the f-value increases, and the p-value corresponding to the f-statistic decreases below 0.05. So, it is acceptable to disregard outliers and important spots.