Ximan Liu
1935858
DSC 423
HW8

I have completed this work independently. The solutions given are entirely my own work.

1)
"All models are wrong, but some are useful" is quite important considering in data science field.
In my understanding, the first thing to pay attention to is to reflect reality accurately and clearly in selecting the data obtained from the observation of the model. Moreover, just burying one's head in a lot of meaningless calculations is not enough to express a good model. In addition, it is not enough to just study the model. We also need to use better interpretation methods to interpret and apply the model. It can be seen from this that it is often not enough to rely solely on computer processing results. What is indispensable is the process of human-computer interaction, such as how to interpret, analyze, and apply data under appropriate circumstances.
We may never be able to perfectly simulate the real behavior or accurate display behind a 100% accurate model, but we can use human-computer interaction and the above operations to be infinitely close to the reality we want to simulate.

2)
a)
For our final model, multicollinearity exists. We can tell that GIR, PABB, PuttingAverage, BB2, SBBB, PuttsPerRound, ADDPA, G2, ADDPPR, PA2 are all greater than 10. They are all necessary.

```
# R code:
m4 <- lm(log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound + G2 + PA2 + BC2 + BB2 +
ADDPA + ADDPPR + DASB + PABB + SBBB, data = d)
summary(m4)
vif(m4)
```

P.S.

```
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
```

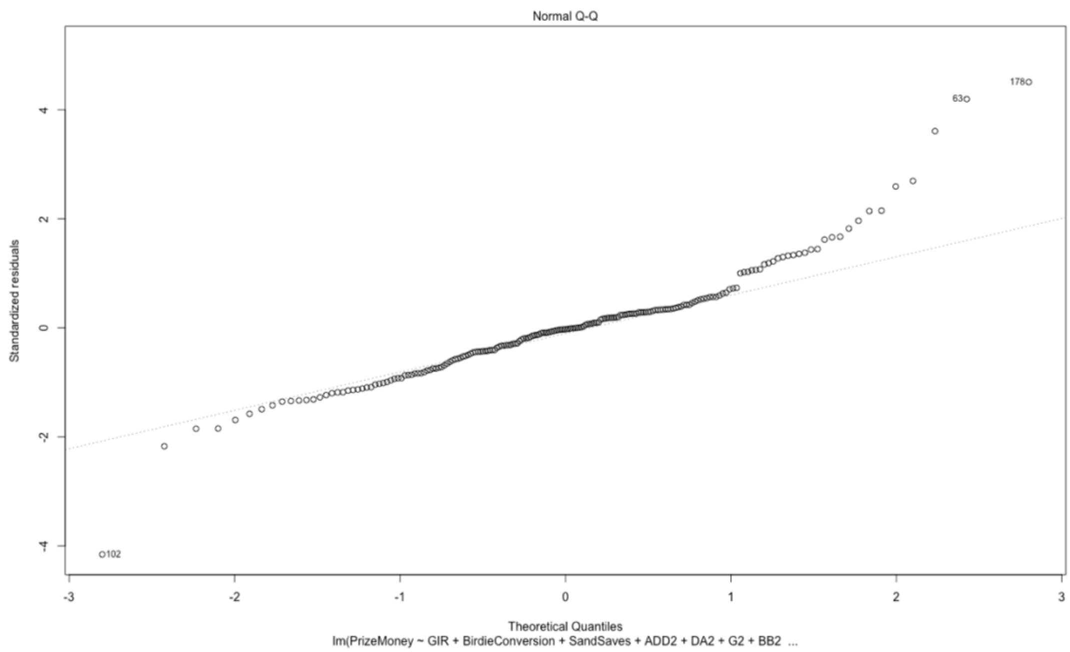
d$DASB <- d$DrivingAccuracy * d$Scrambling

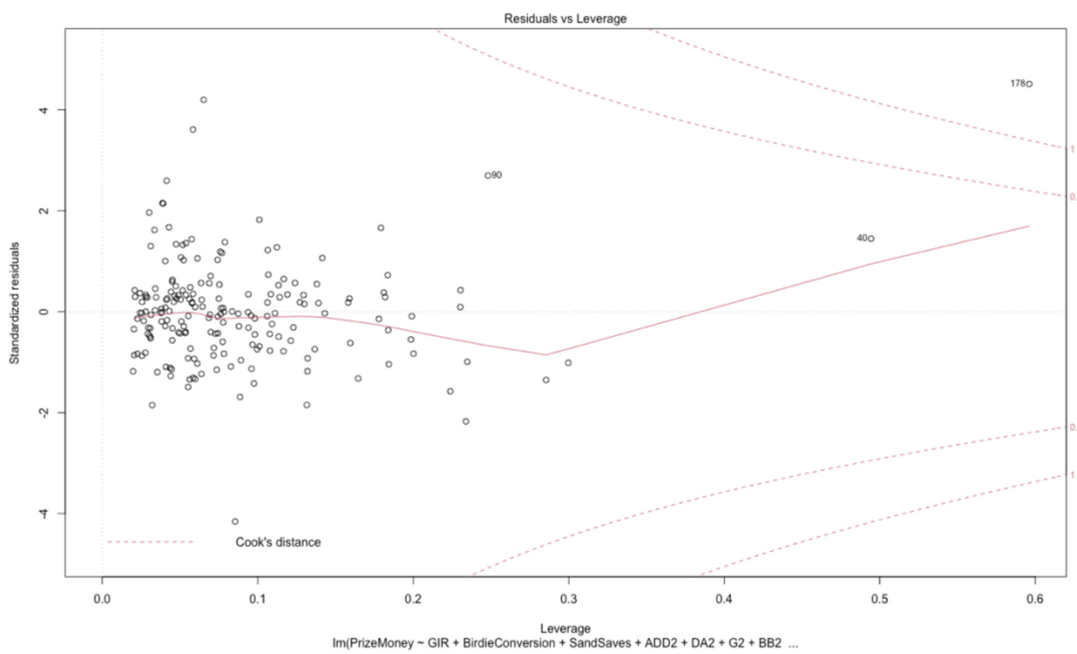
```
> vif(m5)
           GIR PuttingAverage   PuttsPerRound              G2             PA2
    990.930674   15095.465044     2768.725043      997.366537    12471.612566
           BC2            BB2           ADDPA         ADDPPR            DASB
      5.417773     123.283788    14669.159718    17146.356517        5.120463
          PABB           SBBB
    143.318351      41.241249
```

b)
Before:
PrizeMoney = (1.25e+07) + (-2.03e+05) * GIR + (-2.48e+05) * BirdieConversion + (-1.16e+05) * SandSaves + (-4.19e+01) * ADD2 + (2.67e+02) * DA2 + (8.70e+02) * G2 + (7.46e+02) * BB2 + (3.31e+02) * ADDBC + (2.94e+02) * ADDSS + (-5.52e+02) * DAG + (3.47e+04) * GPA + (2.50e+03) * GBC + (-2.26e+04) * PASB + (-4.37e+04) * PABB + (5.60e+02) * SSSB + (8.54e+02) * SBBB + e

P.S.
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

```
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)       1.25e+07   1.21e+06   10.29  < 2e-16 ***
GIR              -2.03e+05   2.97e+04   -6.83  1.3e-10 ***
BirdieConversion -2.48e+05   4.58e+04   -5.41  2.0e-07 ***
SandSaves        -1.16e+05   2.31e+04   -5.03  1.2e-06 ***
ADD2             -4.19e+01   9.14e+00   -4.58  8.7e-06 ***
DA2               2.67e+02   8.54e+01    3.13  0.00202 **
G2                8.70e+02   2.50e+02    3.48  0.00062 ***
BB2               7.46e+02   2.59e+02    2.88  0.00442 **
ADDBC             3.31e+02   1.65e+02    2.00  0.04709 *
ADDSS             2.94e+02   6.29e+01    4.68  5.6e-06 ***
DAG              -5.52e+02   1.69e+02   -3.26  0.00134 **
GPA               3.47e+04   7.52e+03    4.61  7.7e-06 ***
GBC               2.50e+03   4.15e+02    6.02  9.5e-09 ***
PASB             -2.26e+04   4.75e+03   -4.75  4.2e-06 ***
PABB             -4.37e+04   1.12e+04   -3.92  0.00013 ***
SSSB              5.60e+02   1.34e+02    4.19  4.3e-05 ***
SBBB              8.54e+02   3.05e+02    2.80  0.00567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35800 on 179 degrees of freedom
Multiple R-squared:  0.712,     Adjusted R-squared:  0.686
F-statistic: 27.7 on 16 and 179 DF,  p-value: <2e-16
```





Now:

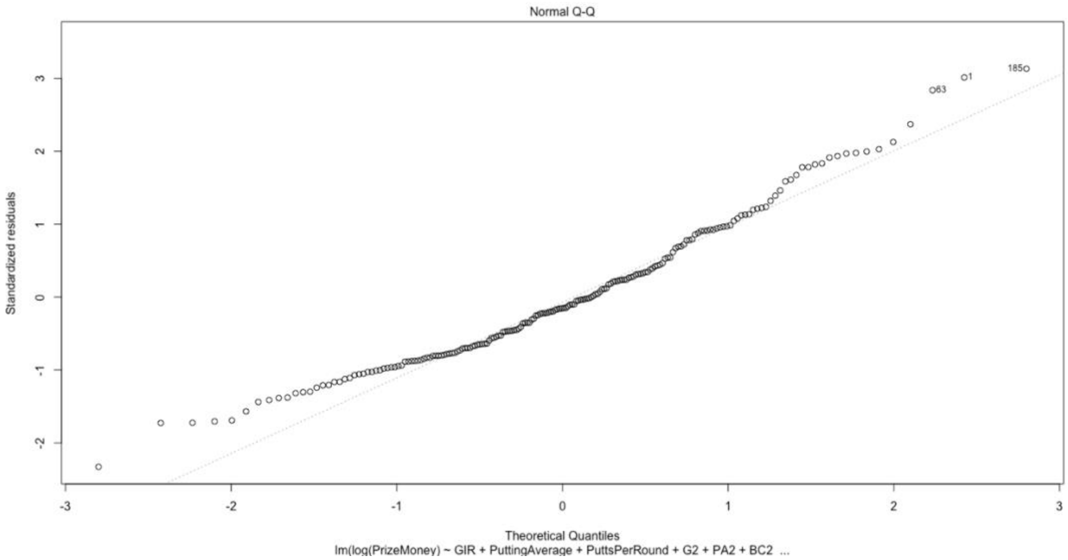
log(PrizeMoney) = (-4.366e+02) + (1.675e+00) * GIR + (6.560e+02) * PuttingAverage + (-1.338e+01) * PuttsPerRound + (-1.132e-02) * G2 + (-1.209e+02) * PA2 + (2.991e-03) * BC2 + (1.085e-02) * BB2 + (-7.591e-01) * ADDPA + (4.573e-02) * ADDPPR + (-5.354e-04) * DASB + (-3.980e-01) * PABB + (5.089e-03) * SBBB + e

P.S.

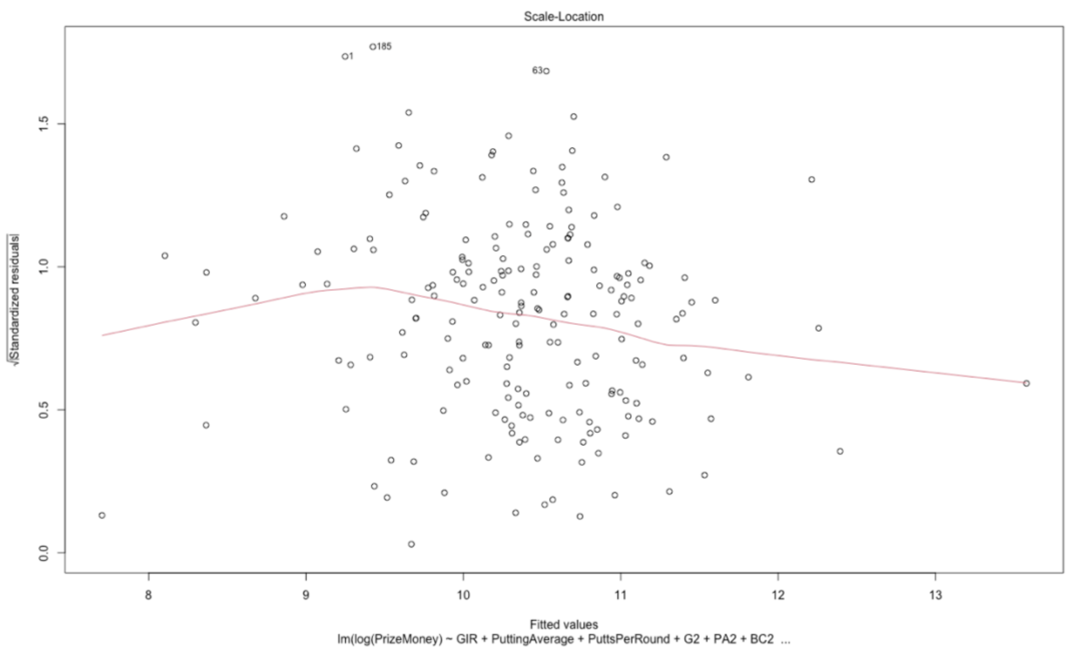
```
d$G2 <- d$GIR^2
d$PA2 <- d$PuttingAverage^2
d$BC2 <- d$BirdieConversion^2
d$PABB <- d$PuttingAverage * d$BounceBack
d$BB2 <- d$BounceBack^2
d$SBBB <- d$Scrambling * d$BounceBack
d$ADDPA <- d$AveDrivingDistance * d$PuttingAverage
d$ADDPPR <- d$AveDrivingDistance * d$PuttsPerRound
d$DASB <- d$DrivingAccuracy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

We can use summary(model) and plot(model) to compare the two models. From the comparison, we can see that the adj-R2 value of the previous model without log transformation is higher. However, the present model with log transform on the normal-QQ graph is closer to a straight diagonal. In that case it represents a linearly line, which means the model is linearly normal distributed. And the corresponding outliers of it are also less. At the same time, the transformation of log function on PrizeMoney (y value) also makes us select different x variables for the final model.

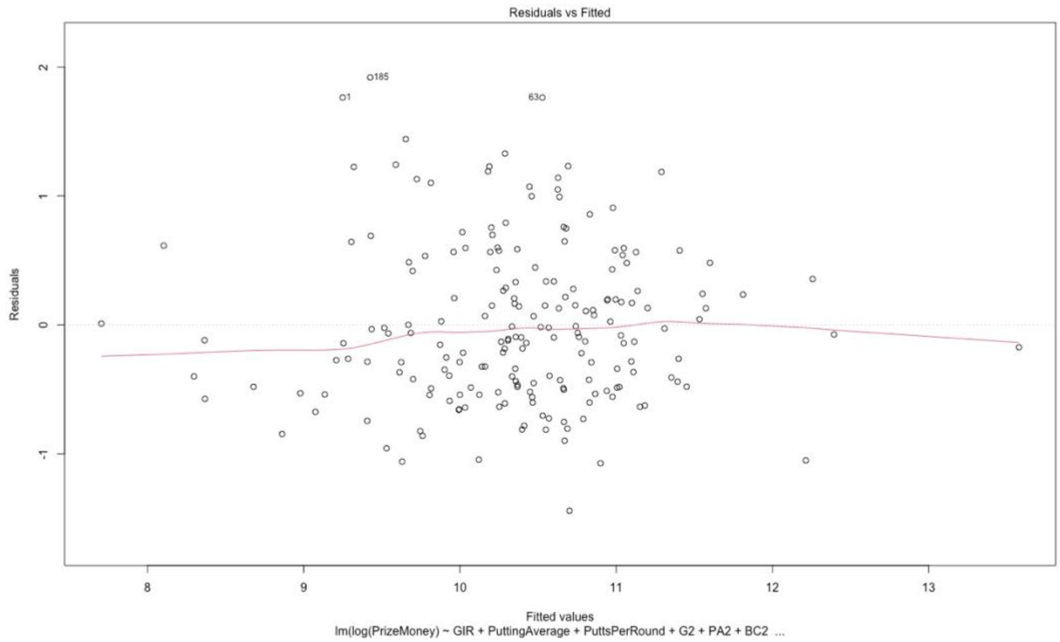
```
Call:
lm(formula = log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound +
    G2 + PA2 + BC2 + BB2 + ADDPA + ADDPPR + DASB + PABB + SBBB,
    data = d)
Residuals:
     Min       1Q   Median       3Q      Max
-1.44164 -0.48010 -0.09378  0.37145  1.91996

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
```

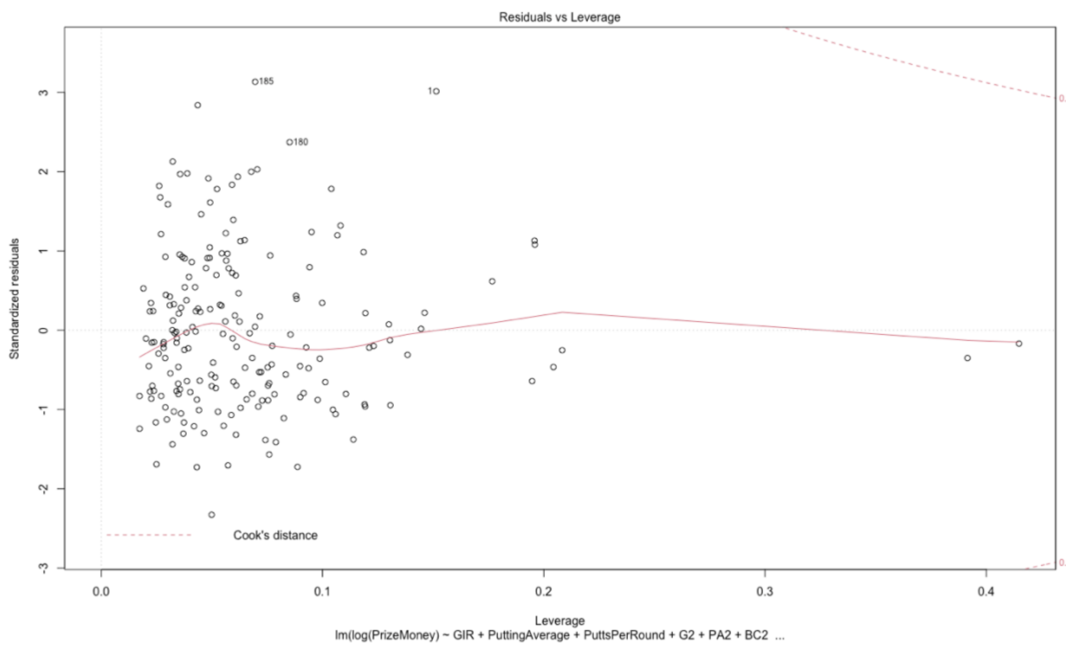


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound + G2 + PA2 + BC2 ...
185
63
1



Scale-Location
√|Standardized residuals|
Fitted values
lm(log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound + G2 + PA2 + BC2 ...
185
1
63





c)

Our present log model of studentized residual graph represents the model is normal distributed with a linear line.

The graphs of x variables are scattered around zero line by random. They represent independency and constant variance.

There are possible outliers in the x variables graphs because studentized residuals are greater than 3 or less than -3.

 # R code:
m4$residuals
sum(m4$residuals)

mean = mean(m4$residuals)
sd = sd(m4$residuals)
resid_zscore = (m4$residuals - mean)/sd

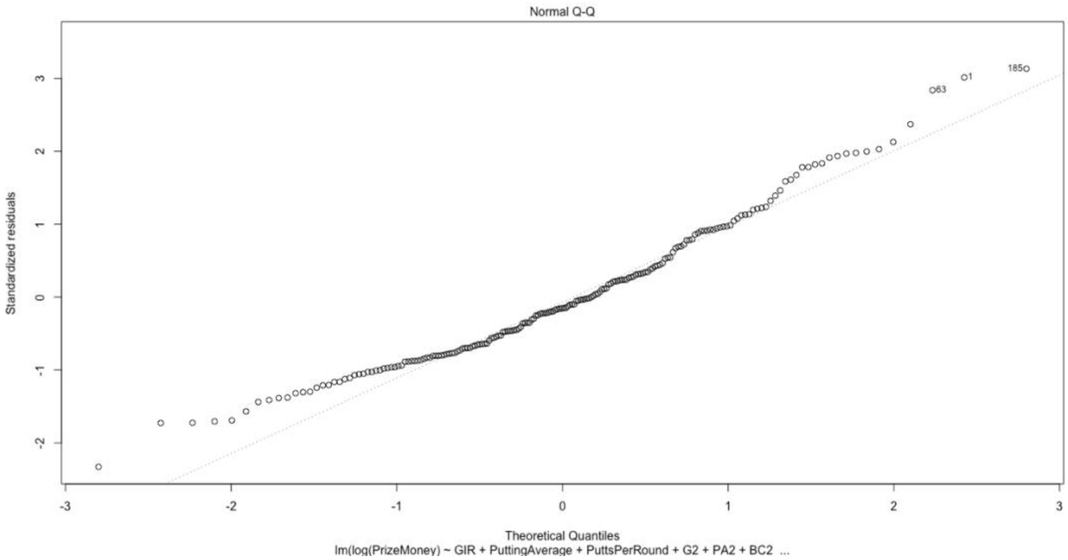
durbinWatsonTest(m4)

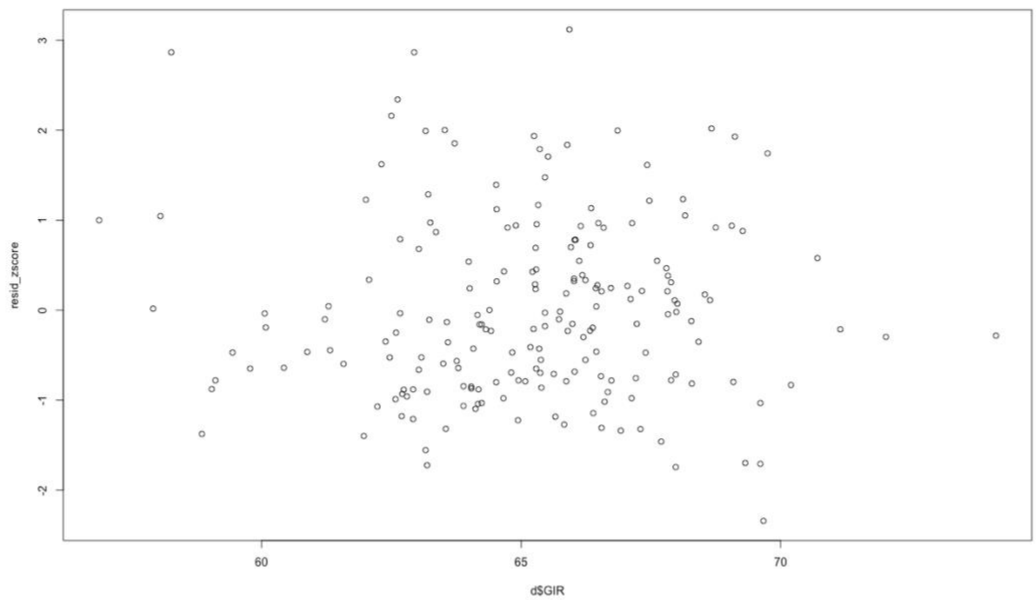
plot(d$GIR, resid_zscore)
plot(d$PuttingAverage, resid_zscore)
plot(d$PuttsPerRound, resid_zscore)
plot(d$G2, resid_zscore)
plot(d$PA2, resid_zscore)
plot(d$BC2, resid_zscore)
plot(d$BB2, resid_zscore)
plot(d$ADDPA, resid_zscore)
plot(d$ADDPPR, resid_zscore)
plot(d$DASB, resid_zscore)
plot(d$PABB, resid_zscore)
plot(d$SBBB, resid_zscore)

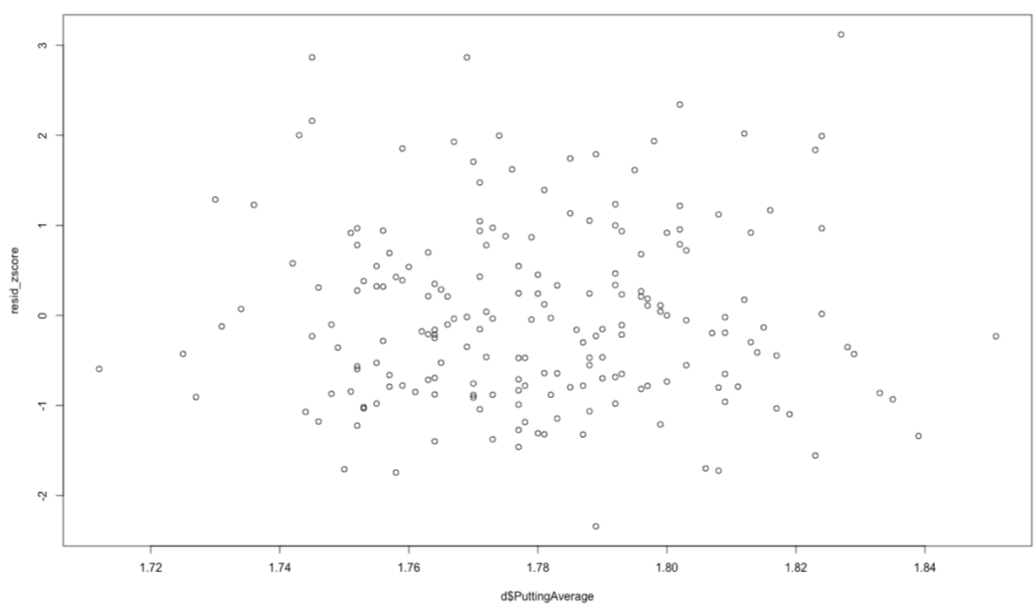
plot(m4)

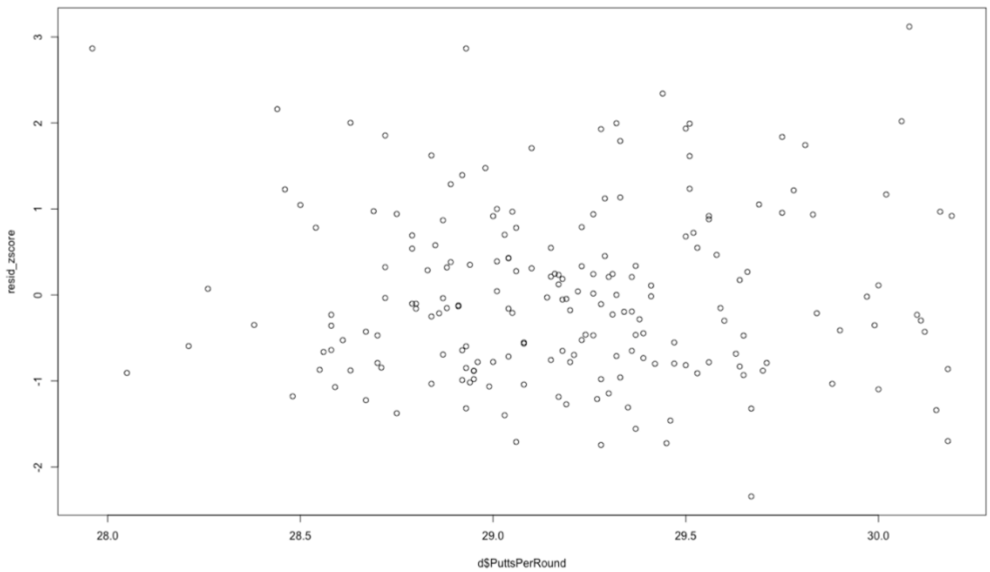
```
> sum(m4$residuals)
[1] -1.720846e-15
```

```
> durbinWatsonTest(m4)
 lag Autocorrelation D-W Statistic p-value
   1      0.07924057      1.778971    0.14
 Alternative hypothesis: rho != 0
```

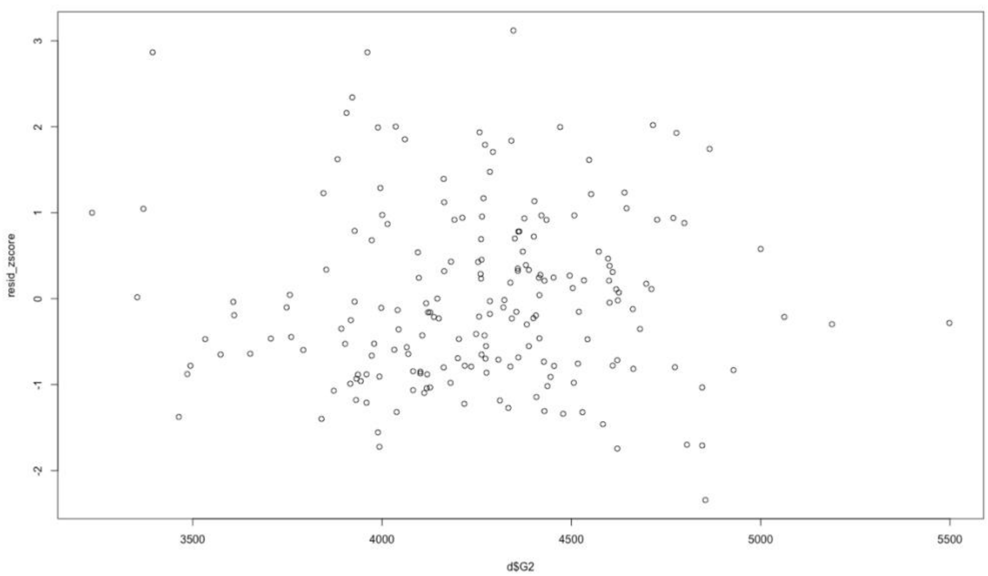

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(PrizeMoney) ~ GIR + PuttingAverage + PuttsPerRound + G2 + PA2 + BC2 ...
185
1
63
-3
-2
-1
0
1
2
3


resid_zscore
3
2
1
0
-1
-2
60
65
70
d$GIR

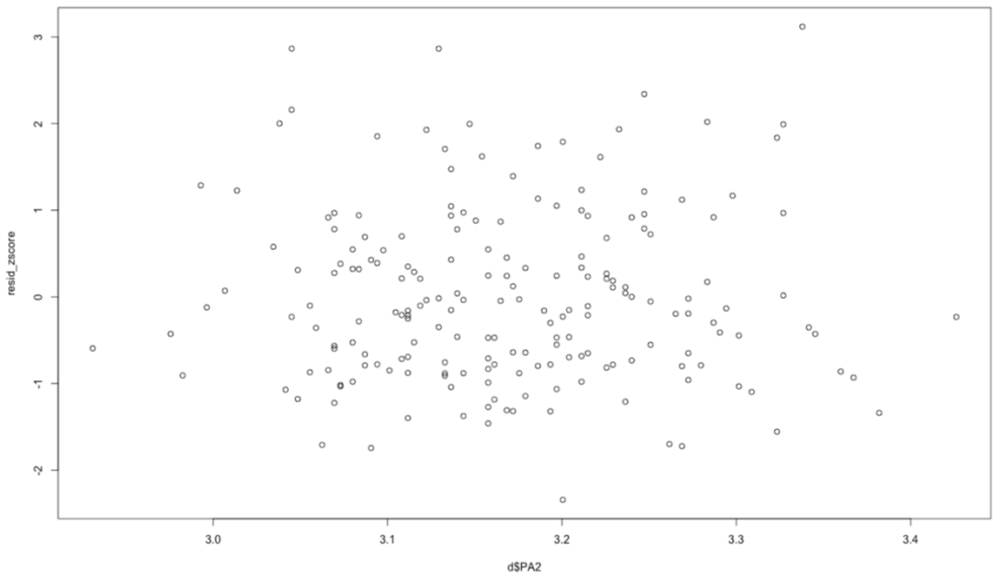

resid_zscore
3
2
1
0
-1
-2
1.72
1.74
1.76
1.78
1.80
1.82
1.84
d$PuttingAverage

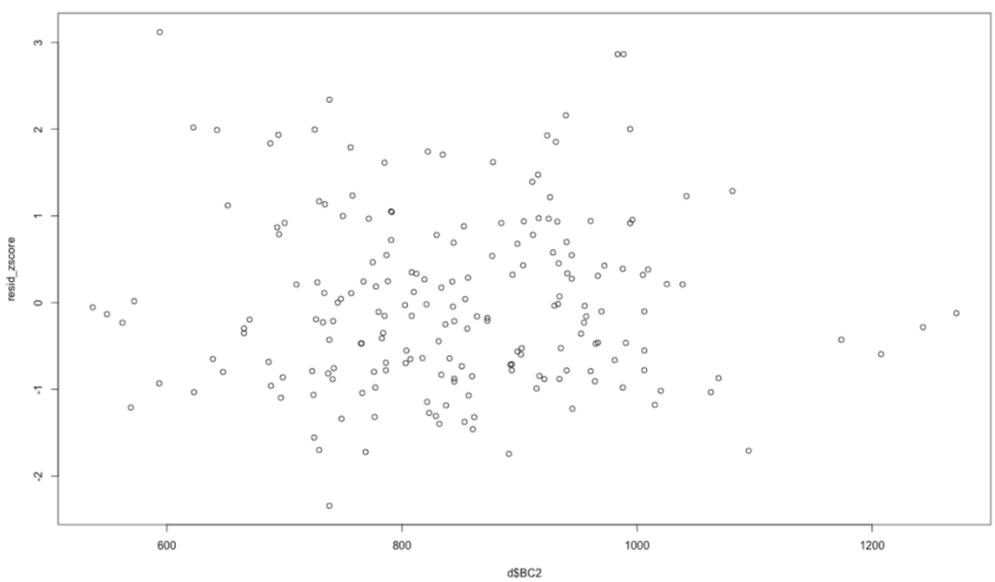


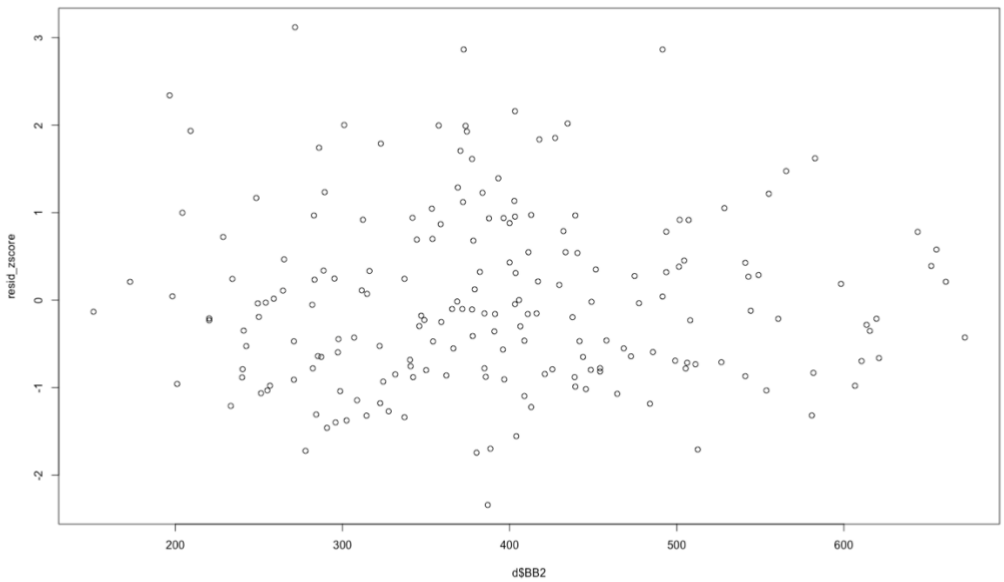
resid_zscore
d$PuttsPerRound

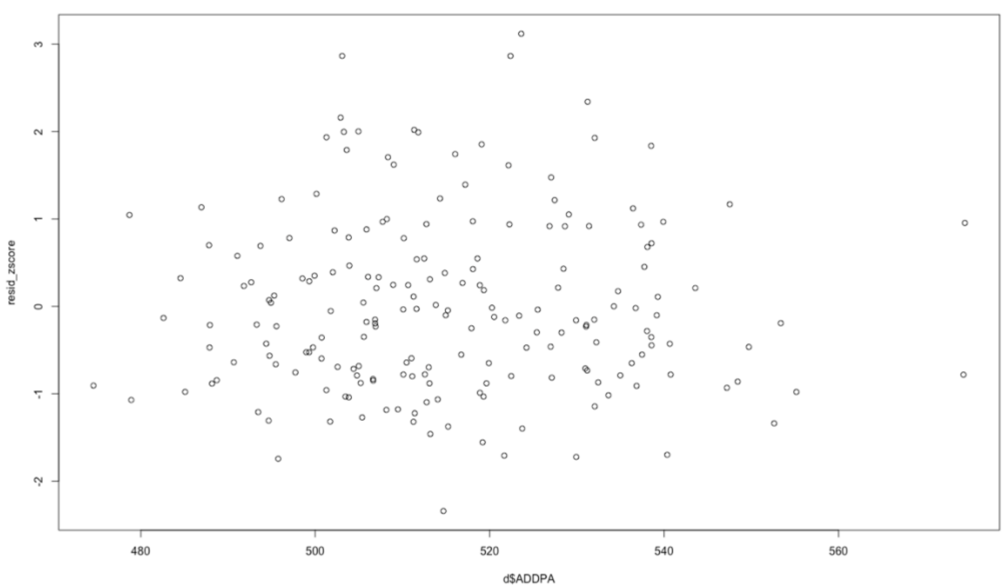


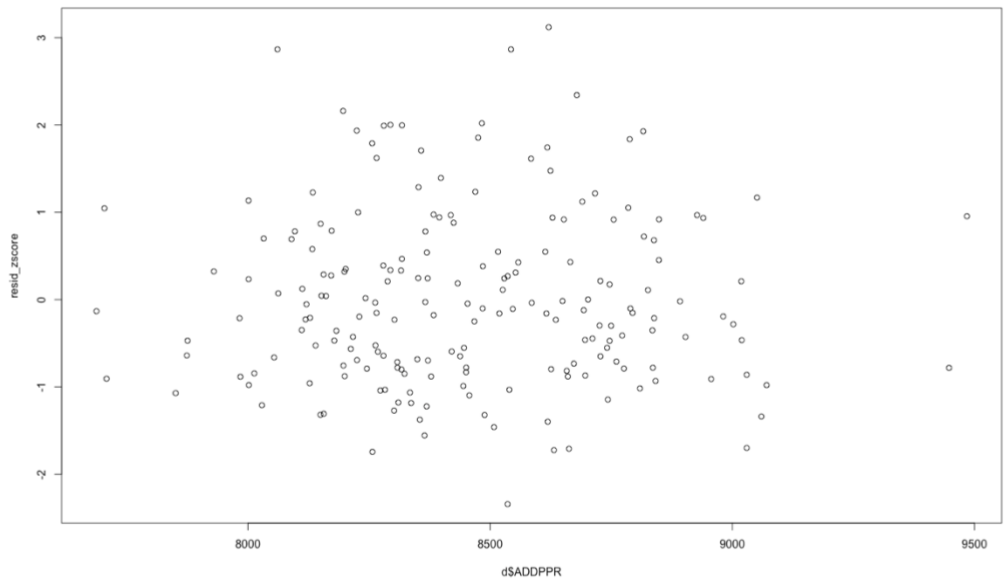
resid_zscore
d$G2


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

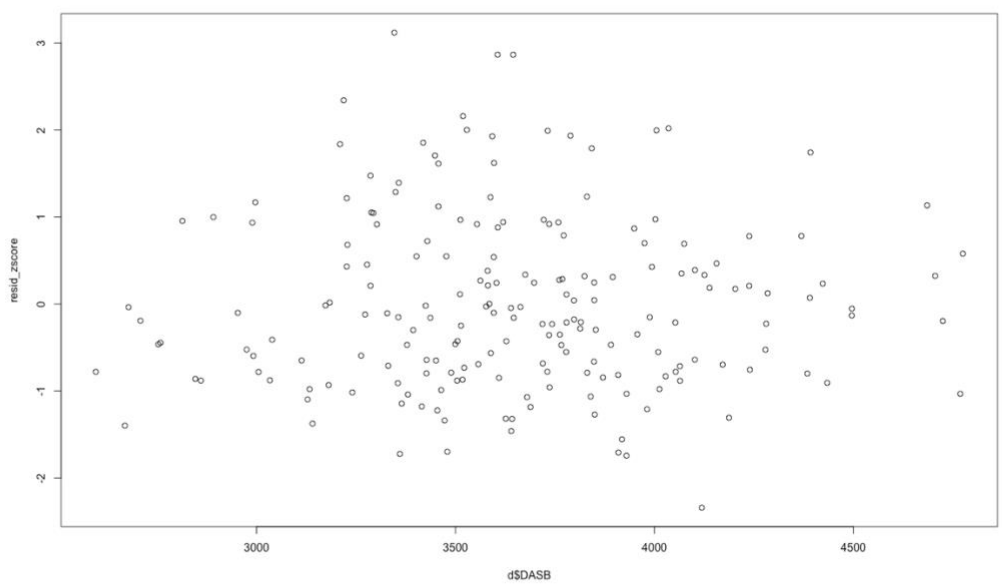

resid_zscore
3
2
1
0
-1
-2
600
800
1000
1200
d$BC2


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

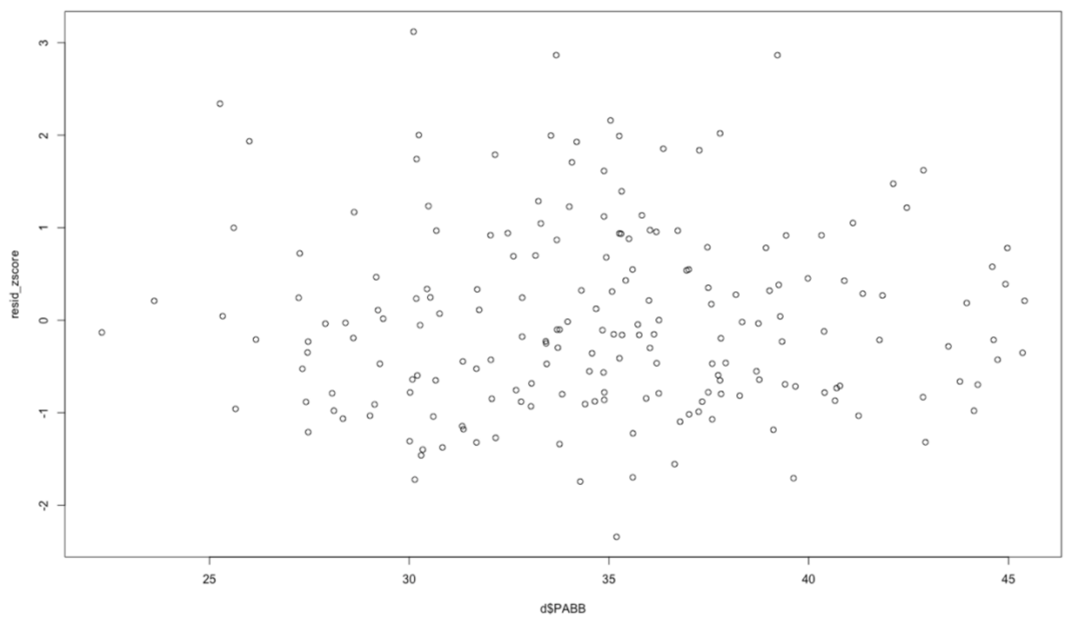

resid_zscore
d$ADDPA
480
500
520
540
560
3
2
1
0
-1
-2

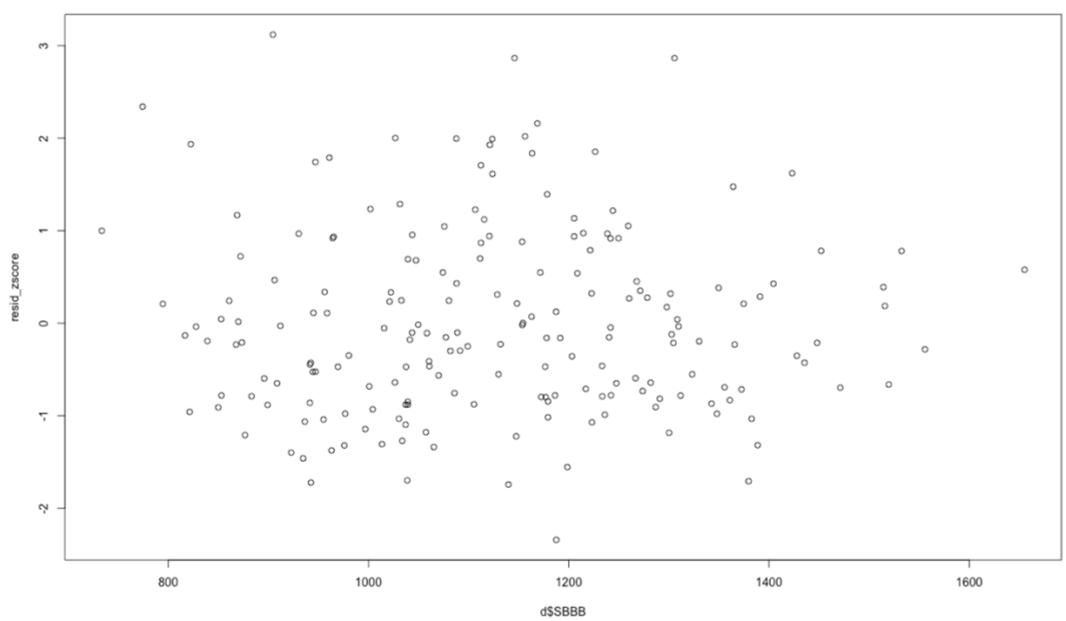


resid_zscore
d$ADDPPR



resid_zscore
d$DASB





d)

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.

There are possible outliers at +3 range in our plots. To deal with that, we can remove the outlier observations 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' observations. 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.