Assignment 6

Biplav Timalsina

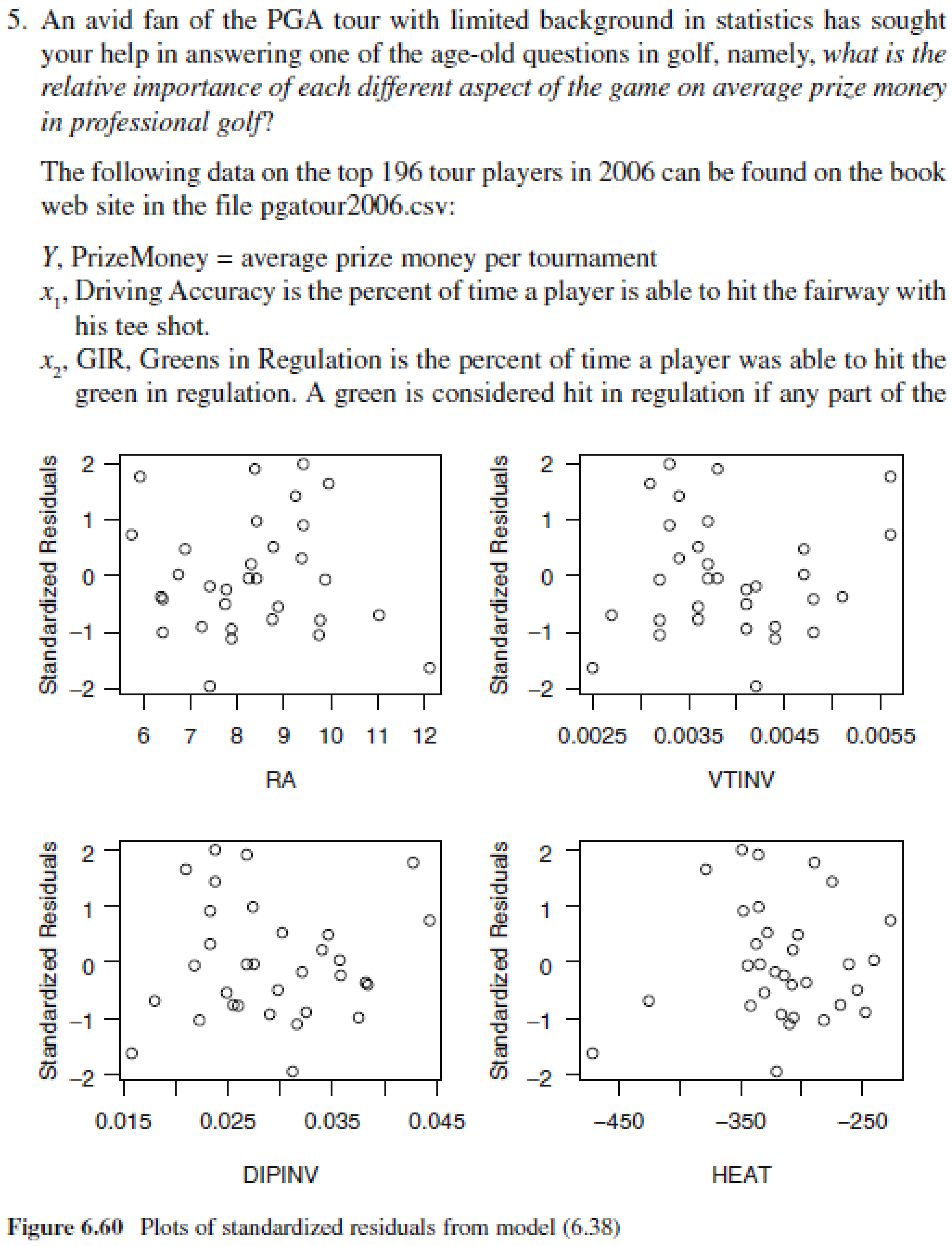
April 08, 2018

STAT 757 Applied Regression Analysis

Instructions [20 points]

Modify this file to provide responses to the Ch.6 Exercises in @sheather2009. You can find some helpful code here: http://www.stat. tamu.edu/~sheather/book/docs/rcode/Chapter6NewMarch2011.R. Also address the project milestones indicated below. Please email **both** your .Rmd (or roxygen .R) and one of the following either .HTML, .PDF, or .DOCX using the format SURNAME-FIRSTNAME-Assignment6.Rmd and SURNAME-FIRSTNAME-Assignment6.pdf.

Exercise 6.7.5 [60 points]

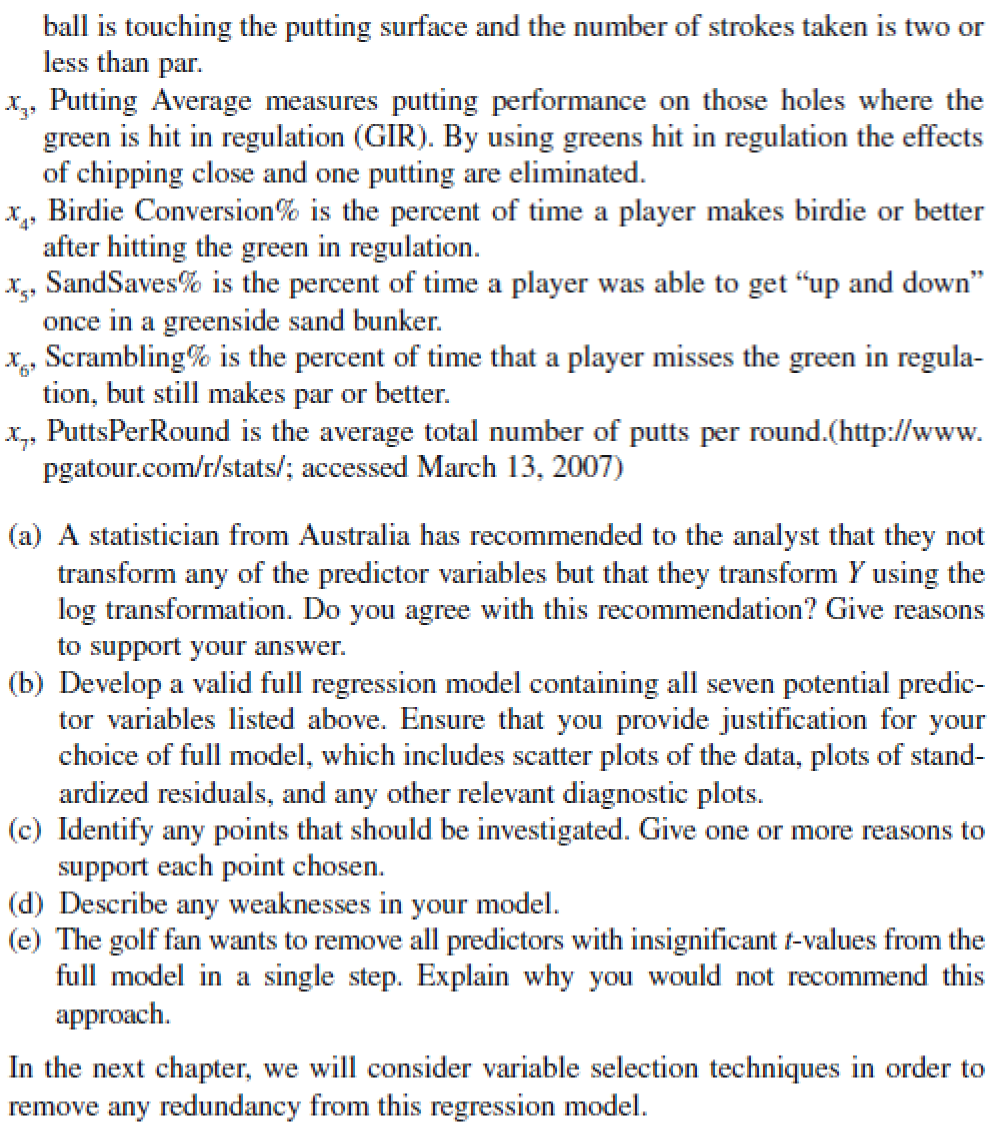


5. An avid fan of the PGA tour with limited background in statistics has sought your help in answering one of the age-old questions in golf, namely, *what is the relative importance of each different aspect of the game on average prize money in professional golf?*

   The following data on the top 196 tour players in 2006 can be found on the book web site in the file pgatour2006.csv:

   Y, PrizeMoney = average prize money per tournament
   x_1, Driving Accuracy is the percent of time a player is able to hit the fairway with his tee shot.
   x_2, GIR, Greens in Regulation is the percent of time a player was able to hit the green in regulation. A green is considered hit in regulation if any part of the

Figure 6.60 Plots of standardized residuals from model (6.38)



ball is touching the putting surface and the number of strokes taken is two or less than par.

x_3, Putting Average measures putting performance on those holes where the green is hit in regulation (GIR). By using greens hit in regulation the effects of chipping close and one putting are eliminated.

x_4, Birdie Conversion% is the percent of time a player makes birdie or better after hitting the green in regulation.

x_5, SandSaves% is the percent of time a player was able to get “up and down” once in a greenside sand bunker.

x_6, Scrambling% is the percent of time that a player misses the green in regulation, but still makes par or better.

x_7, PuttsPerRound is the average total number of putts per round.(http://www.pgatour.com/r/stats/; accessed March 13, 2007)

(a) A statistician from Australia has recommended to the analyst that they not transform any of the predictor variables but that they transform Y using the log transformation. Do you agree with this recommendation? Give reasons to support your answer.

(b) Develop a valid full regression model containing all seven potential predictor variables listed above. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots.

(c) Identify any points that should be investigated. Give one or more reasons to support each point chosen.

(d) Describe any weaknesses in your model.

(e) The golf fan wants to remove all predictors with insignificant t-values from the full model in a single step. Explain why you would not recommend this approach.

In the next chapter, we will consider variable selection techniques in order to remove any redundancy from this regression model.

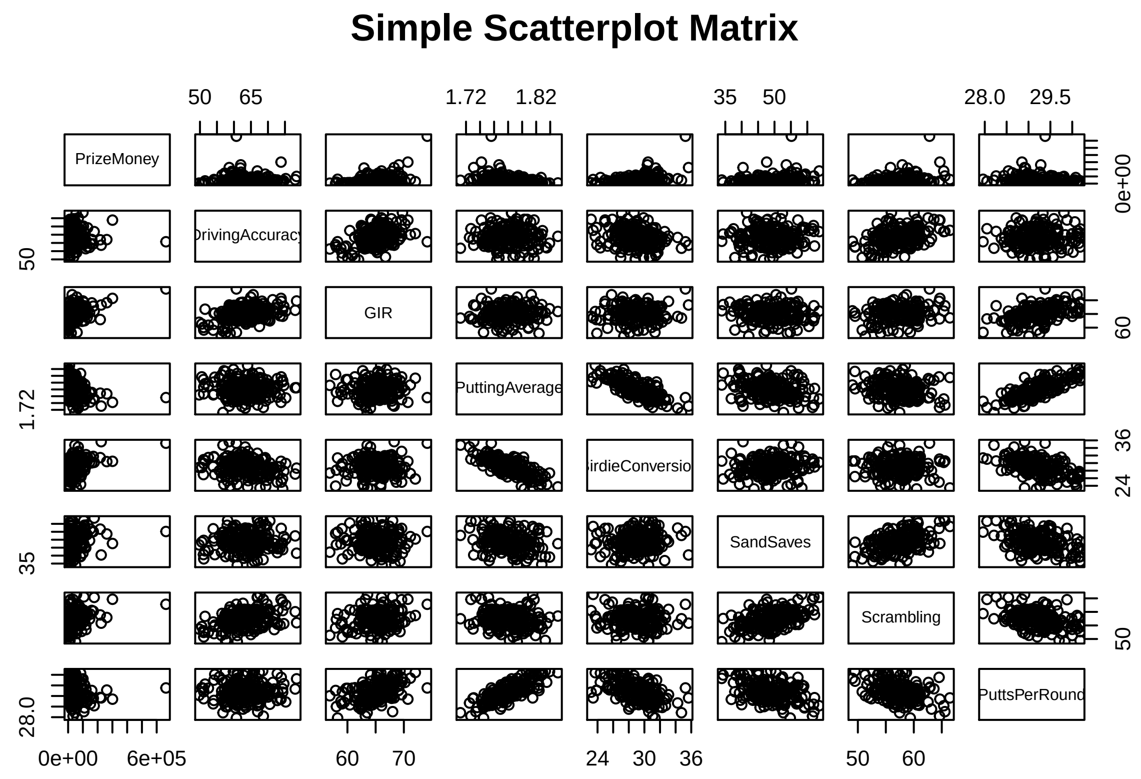
Let us first plot the scatterplot of the different variables in our data to have a general idea about our data.

```
PGA<-read.csv("F:/unr/4th sem/applied regression analysis/Assignments/HW6/pgatour2006.csv",header = TRU
attach(PGA)
newPGA<-PGA[,c(3,4,5,6,7,8,9,10,11,12)]

cor(newPGA)
```

```
##                     PrizeMoney AveDrivingDistance DrivingAccuracy
## PrizeMoney          1.00000000         0.15900129     0.024677039
## AveDrivingDistance  0.15900129         1.00000000    -0.590599303
## DrivingAccuracy     0.02467704        -0.59059930     1.000000000
## GIR                 0.41021935         0.16460354     0.416356043
## PuttingAverage     -0.31305150         0.08595947    -0.025582688
## BirdieConversion    0.41342953         0.37568272    -0.252125225
## SandSaves           0.22187452        -0.23669494     0.035407734
## Scrambling          0.28472059        -0.38033753     0.396059676
## BounceBack          0.33620030         0.23750860     0.001746659
## PuttsPerRound      -0.11249143         0.25656855     0.060313852
##                             GIR PuttingAverage BirdieConversion   SandSaves
## PrizeMoney          0.41021935    -0.31305150       0.41342953  0.22187452
## AveDrivingDistance  0.16460354     0.08595947       0.37568272 -0.23669494
## DrivingAccuracy     0.41635604    -0.02558269      -0.25212523  0.03540773
## GIR                 1.00000000     0.05880737       0.02685014 -0.08107691
## PuttingAverage      0.05880737     1.00000000      -0.76795939 -0.26509212
## BirdieConversion    0.02685014    -0.76795939       1.00000000  0.13050563
## SandSaves          -0.08107691    -0.26509212       0.13050563  1.00000000
## Scrambling          0.19435094    -0.19894278      -0.02891616  0.49626530
## BounceBack          0.29275929    -0.31856723       0.48262514  0.02628417
## PuttsPerRound       0.48083985     0.79168281      -0.50072564 -0.42046979
##                     Scrambling   BounceBack PuttsPerRound
## PrizeMoney          0.28472059  0.336200304   -0.11249143
## AveDrivingDistance -0.38033753  0.237508596    0.25656855
## DrivingAccuracy     0.39605968  0.001746659    0.06031385
## GIR                 0.19435094  0.292759294    0.48083985
## PuttingAverage     -0.19894278 -0.318567233    0.79168281
## BirdieConversion   -0.02891616  0.482625137   -0.50072564
## SandSaves           0.49626530  0.026284174   -0.42046979
## Scrambling          1.00000000  0.087693355   -0.41558527
## BounceBack          0.08769336  1.000000000   -0.13501883
## PuttsPerRound      -0.41558527 -0.135018830    1.00000000
```

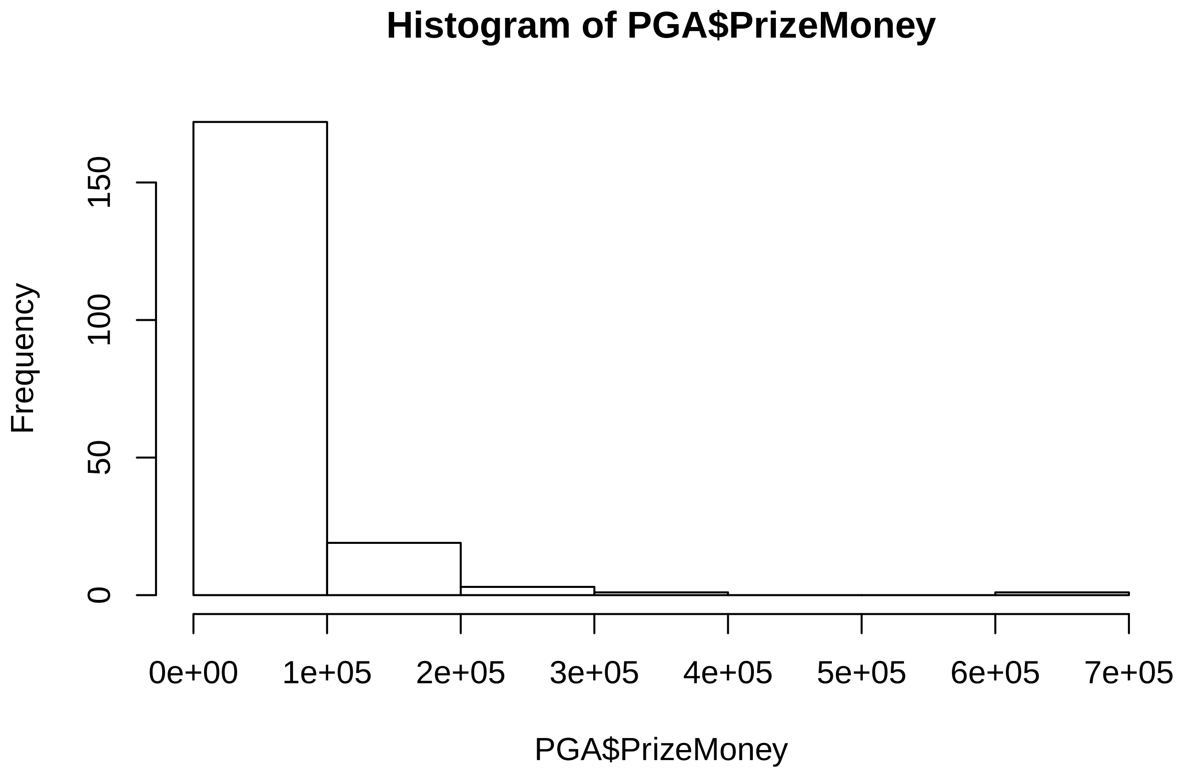
```
#pairs(newData[,1:5])
pairs(~PrizeMoney
+DrivingAccuracy
+GIR
+PuttingAverage
+BirdieConversion
+SandSaves
+Scrambling
+PuttsPerRound
, data=newPGA,main="Simple Scatterplot Matrix")
```



a) Ans:

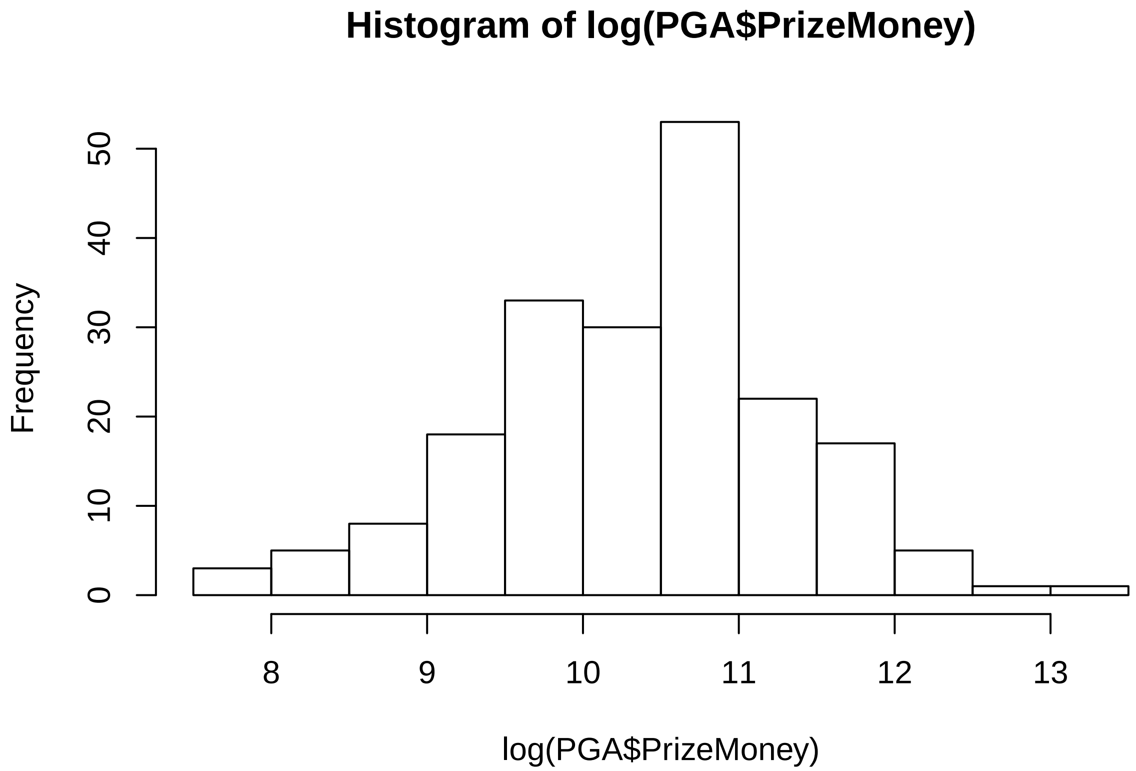
Let us plot the histogram for the PrizeMoney data points to observe how it is distributed.

```
hist(PGA$PrizeMoney)
```



It seems that the PrizeMoney data is left skewed as can be seen above. If we apply log transformation to this dependent variable, can we make it normally distributed? Lets try.

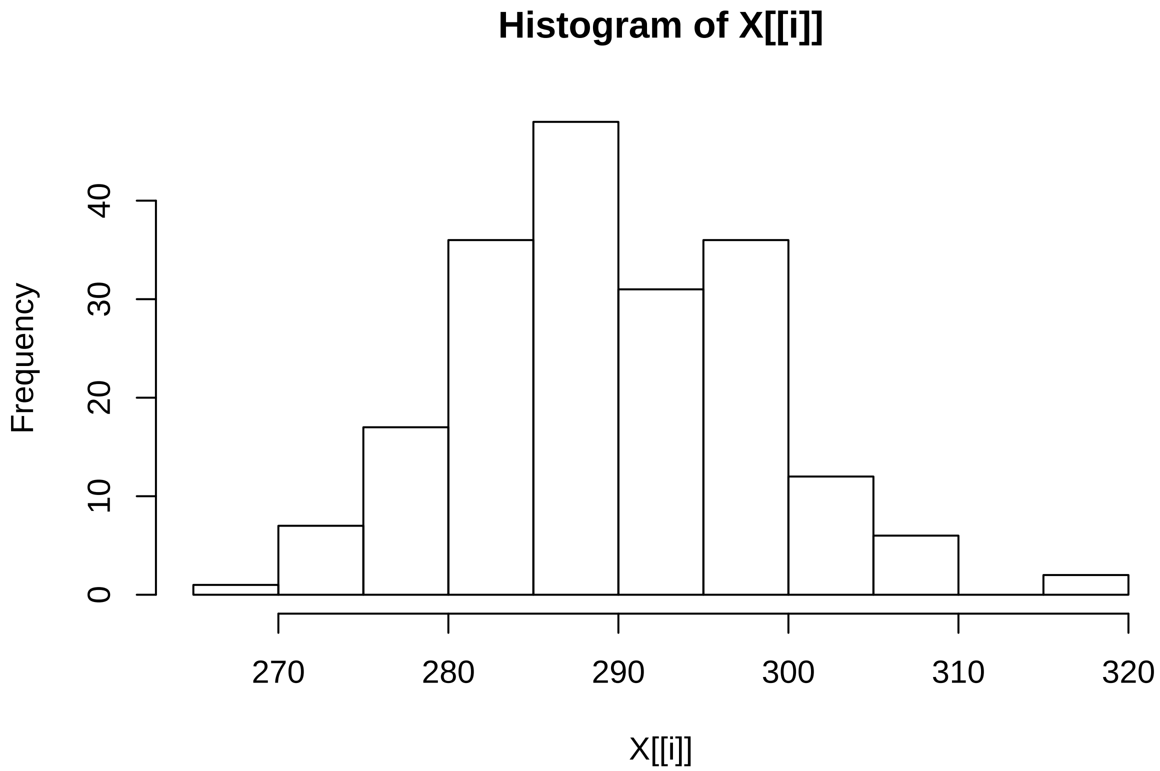
```
hist(log(PGA$PrizeMoney))
```

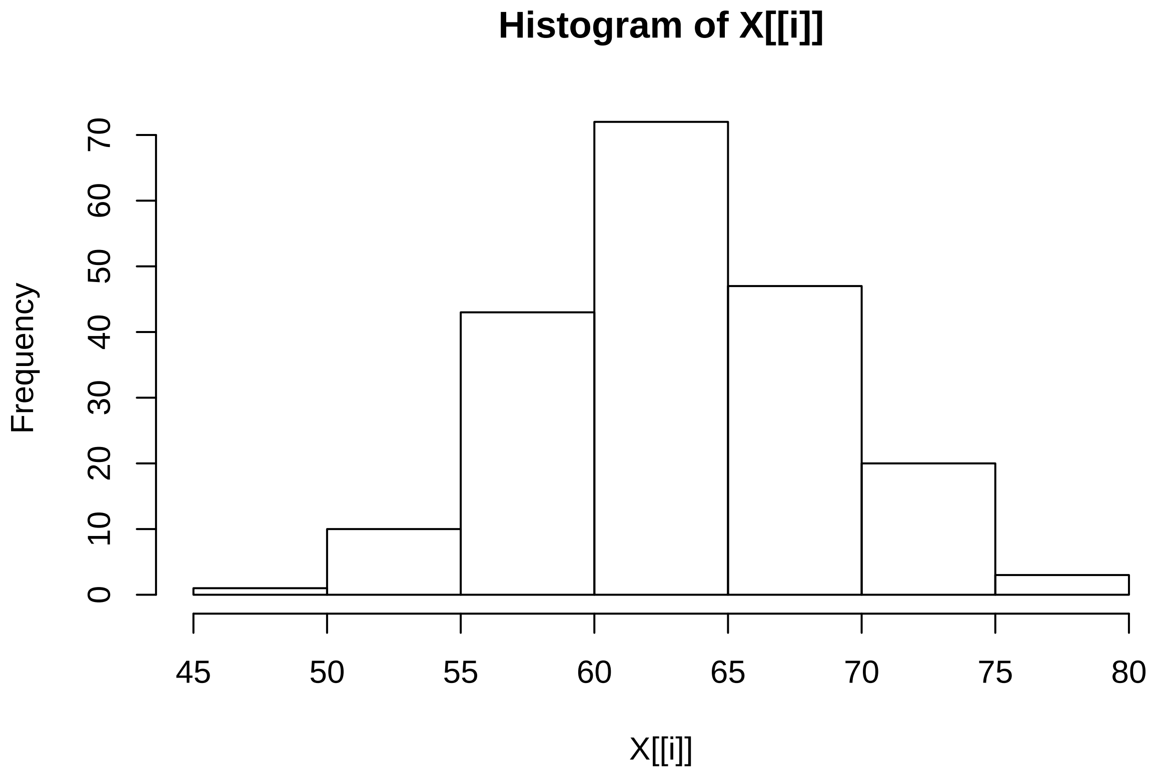


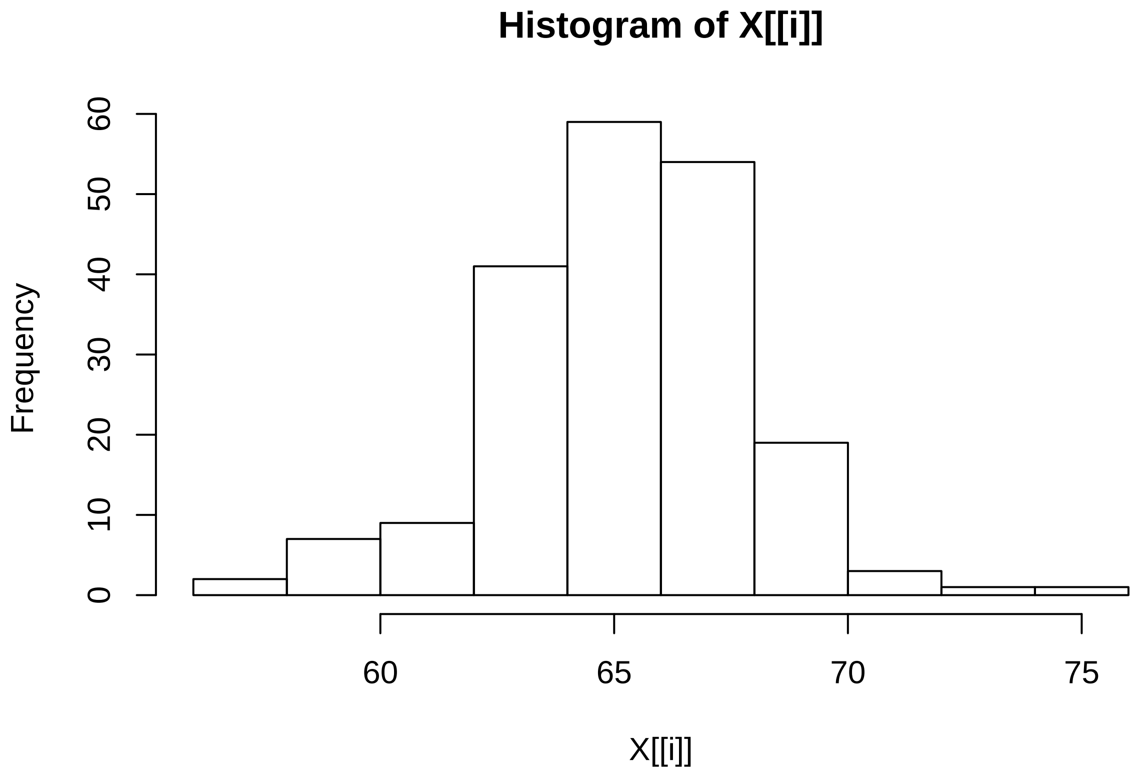
Now, it seems normally distributed. Thus, we should definitely log transform our PrizeMoney (y) variable.

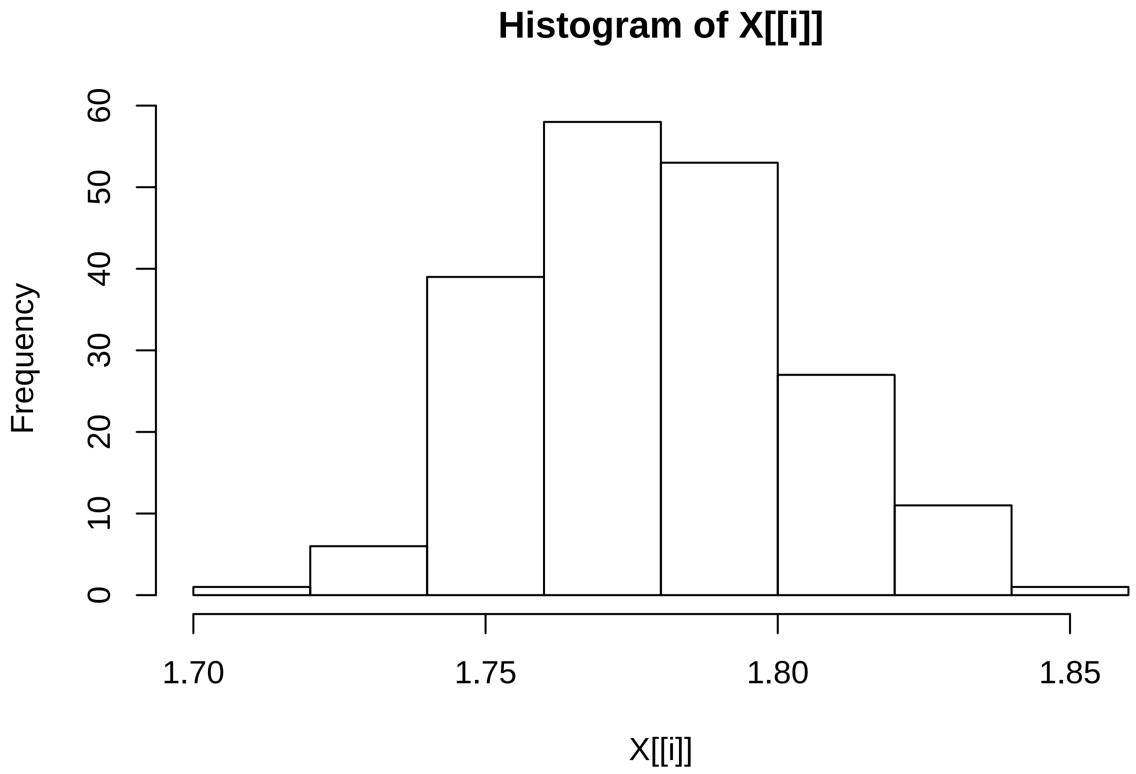
What about other x variables? The statistician has recommended to not perform any transformation the variables. Lets observe the histogram for all the predictors.

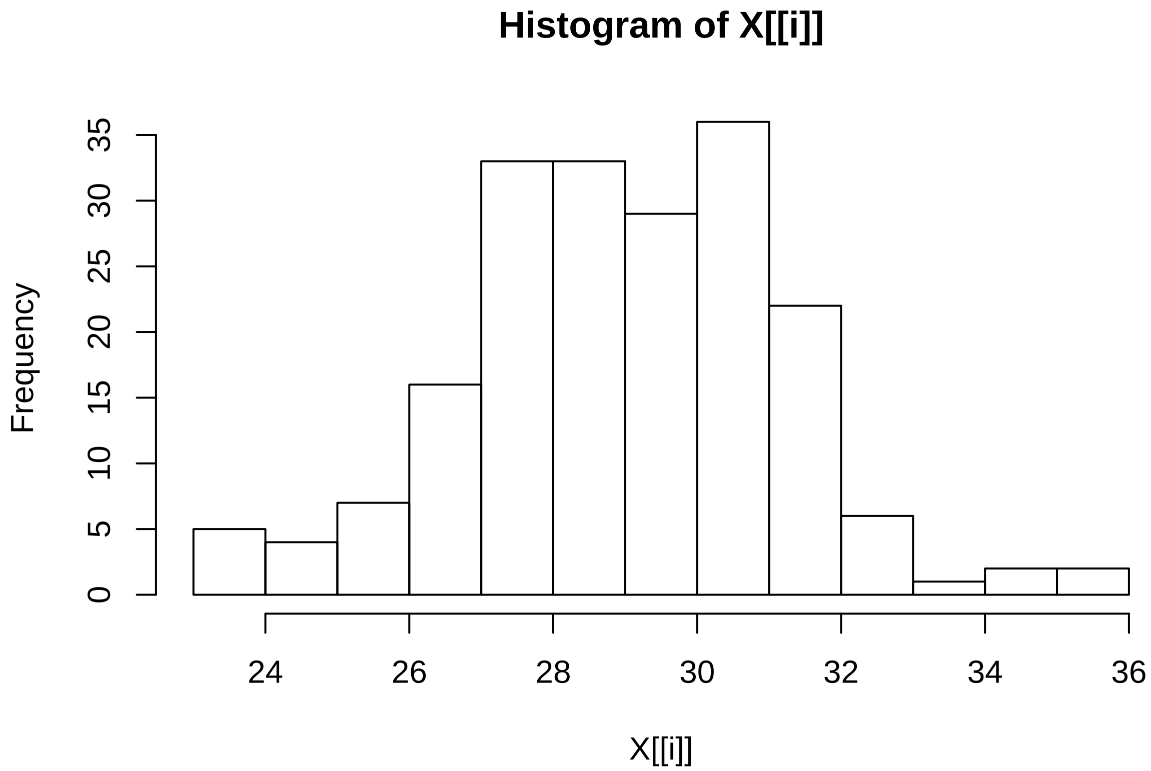
```
lapply(PGA[4:12],FUN=hist)
```

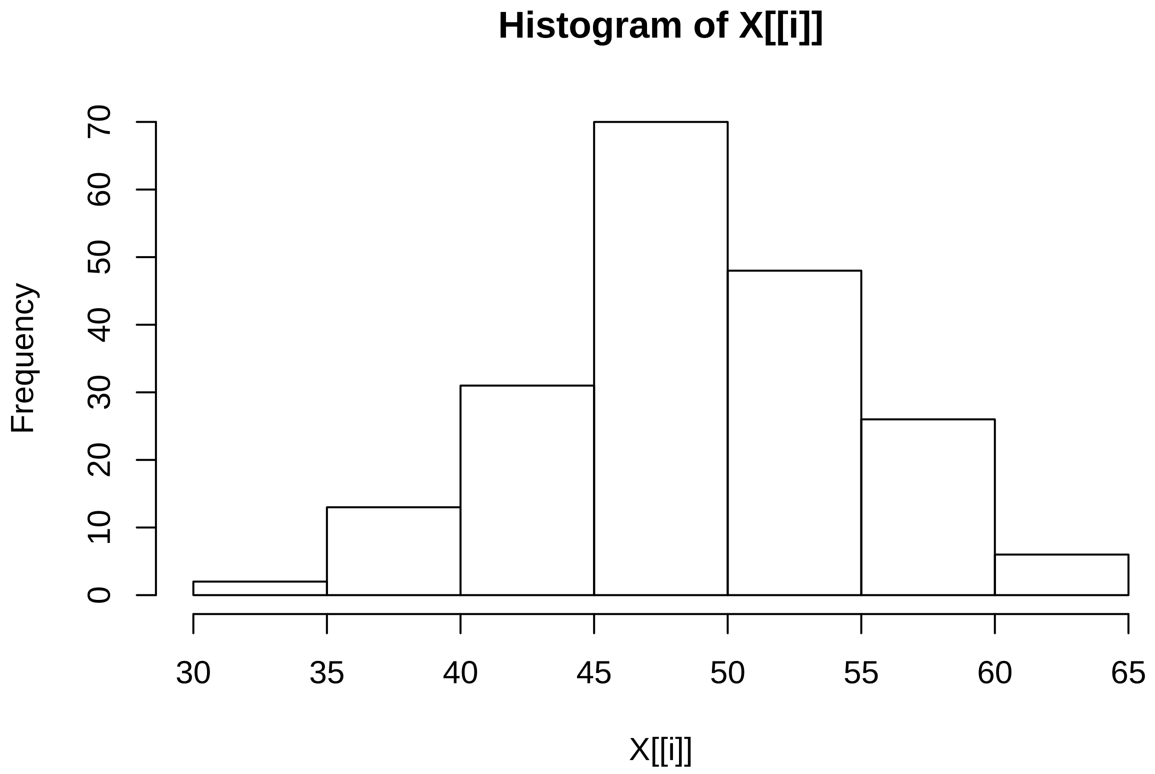

Histogram of X[[i]]
Frequency
0
10
20
30
40
270
280
290
300
310
320
X[[i]]

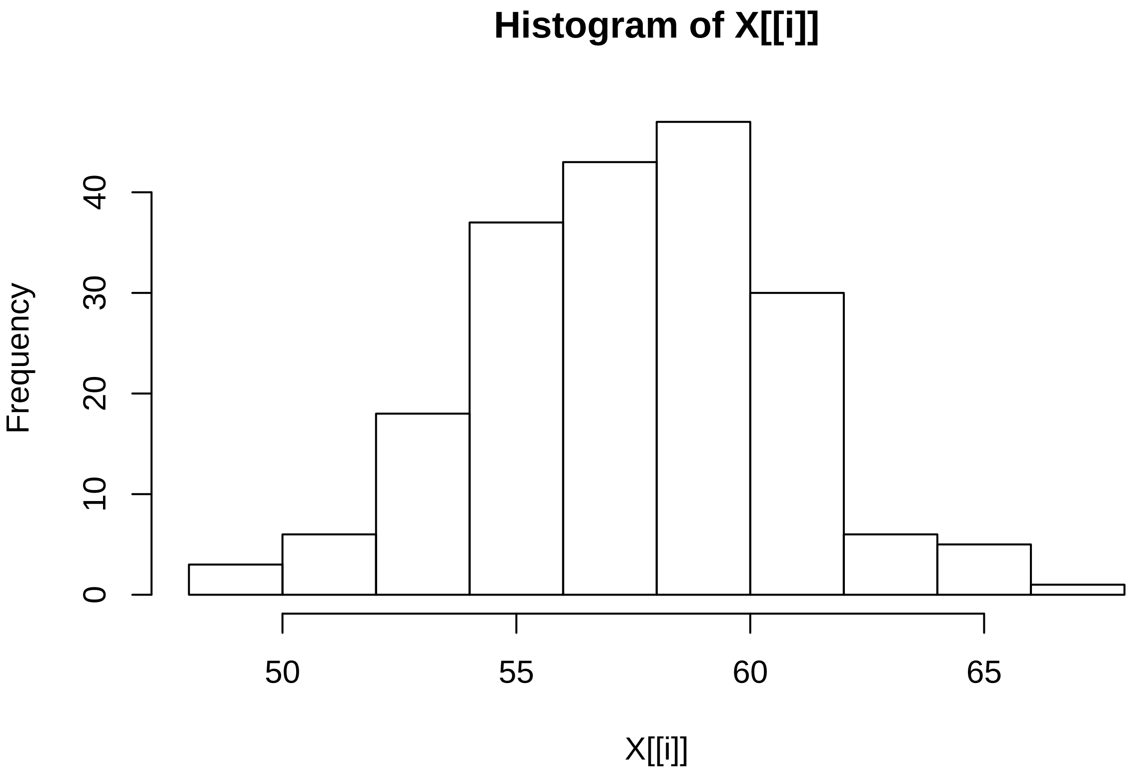

Histogram of X[[i]]
Frequency
0
10
20
30
40
50
60
70
45
50
55
60
65
70
75
80
X[[i]]

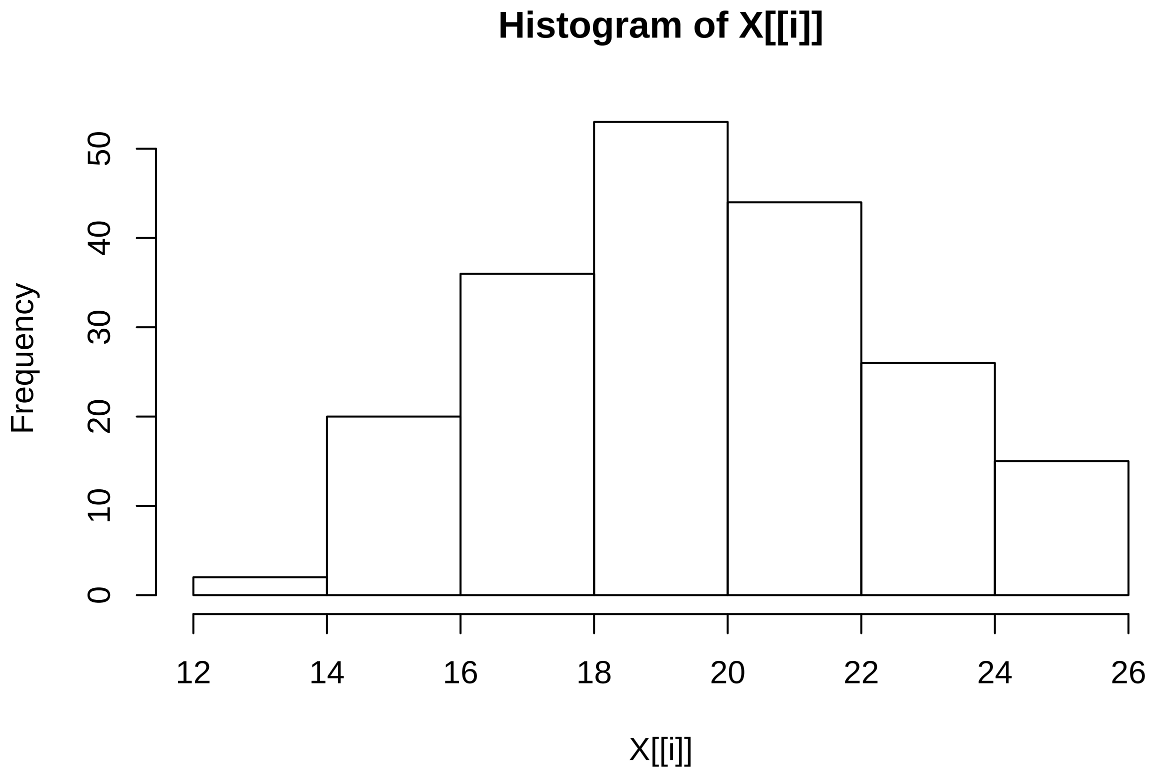

Histogram of X[[i]]
Frequency
0
10
20
30
40
50
60
60
65
70
75
X[[i]]

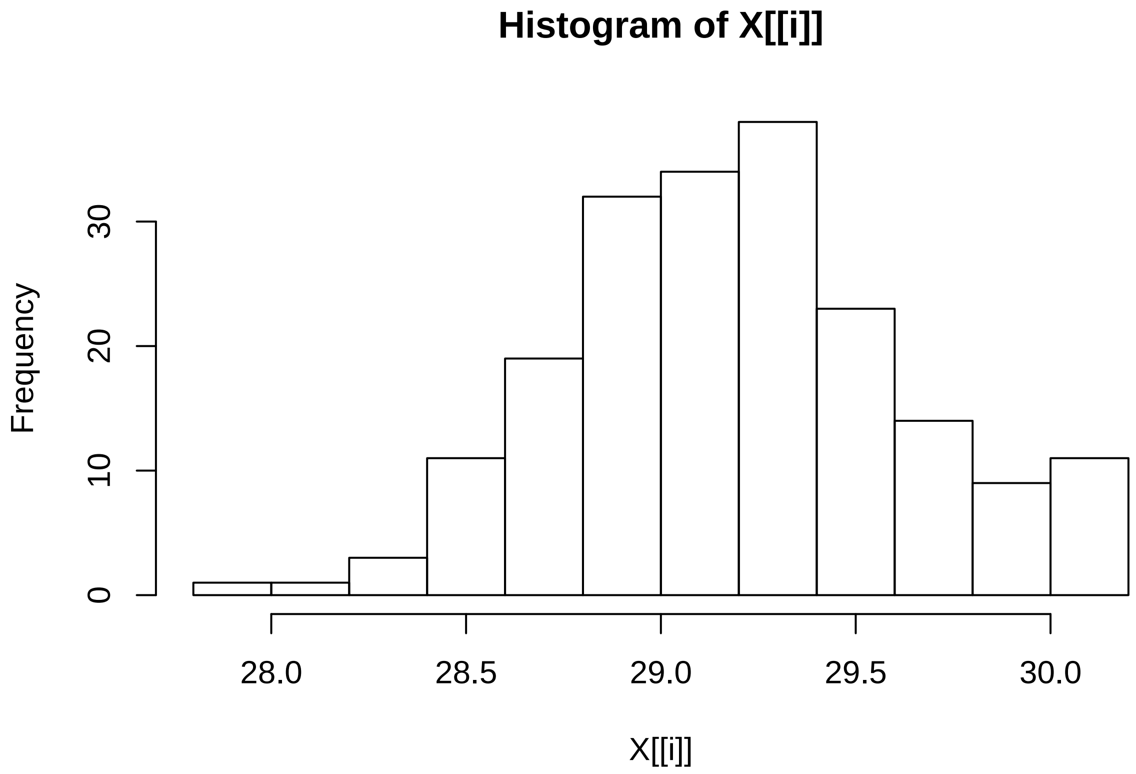

Histogram of X[[i]]
Frequency
0
10
20
30
40
50
60
1.70
1.75
1.80
1.85
X[[i]]


Histogram of X[[i]]
Frequency
0
5
10
15
20
25
30
35
24
26
28
30
32
34
36
X[[i]]


Histogram of X[[i]]
Frequency
0
10
20
30
40
50
60
70
30
35
40
45
50
55
60
65
X[[i]]


Histogram of X[[i]]
Frequency
0
10
20
30
40
50
55
60
65
X[[i]]


Histogram of X[[i]]
Frequency
0
10
20
30
40
50
12
14
16
18
20
22
24
26
X[[i]]



```
## $AveDrivingDistance
## $breaks
##  [1] 265 270 275 280 285 290 295 300 305 310 315 320
##
## $counts
##  [1]  1  7 17 36 48 31 36 12  6  0  2
##
## $density
##  [1] 0.001020408 0.007142857 0.017346939 0.036734694 0.048979592
##  [6] 0.031632653 0.036734694 0.012244898 0.006122449 0.000000000
## [11] 0.002040816
##
## $mids
##  [1] 267.5 272.5 277.5 282.5 287.5 292.5 297.5 302.5 307.5 312.5 317.5
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $DrivingAccuracy
## $breaks
## [1] 45 50 55 60 65 70 75 80
##
## $counts
## [1]  1 10 43 72 47 20  3
##
## $density
## [1] 0.001020408 0.010204082 0.043877551 0.073469388 0.047959184 0.020408163
## [7] 0.003061224
##
## $mids
## [1] 47.5 52.5 57.5 62.5 67.5 72.5 77.5
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $GIR
## $breaks
##  [1] 56 58 60 62 64 66 68 70 72 74 76
##
## $counts
##  [1]  2  7  9 41 59 54 19  3  1  1
##
## $density
##  [1] 0.005102041 0.017857143 0.022959184 0.104591837 0.150510204
##  [6] 0.137755102 0.048469388 0.007653061 0.002551020 0.002551020
##
## $mids
##  [1] 57 59 61 63 65 67 69 71 73 75
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $PuttingAverage
## $breaks
## [1] 1.70 1.72 1.74 1.76 1.78 1.80 1.82 1.84 1.86
##
## $counts
## [1]  1  6 39 58 53 27 11  1
##
## $density
## [1]  0.255102  1.530612  9.948980 14.795918 13.520408  6.887755  2.806122
## [8]  0.255102
```

```
##
## $mids
## [1] 1.71 1.73 1.75 1.77 1.79 1.81 1.83 1.85
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $BirdieConversion
## $breaks
##  [1] 23 24 25 26 27 28 29 30 31 32 33 34 35 36
##
## $counts
##  [1]  5  4  7 16 33 33 29 36 22  6  1  2  2
##
## $density
##  [1] 0.025510204 0.020408163 0.035714286 0.081632653 0.168367347
##  [6] 0.168367347 0.147959184 0.183673469 0.112244898 0.030612245
## [11] 0.005102041 0.010204082 0.010204082
##
## $mids
##  [1] 23.5 24.5 25.5 26.5 27.5 28.5 29.5 30.5 31.5 32.5 33.5 34.5 35.5
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $SandSaves
## $breaks
## [1] 30 35 40 45 50 55 60 65
##
## $counts
## [1]  2 13 31 70 48 26  6
##
## $density
## [1] 0.002040816 0.013265306 0.031632653 0.071428571 0.048979592 0.026530612
## [7] 0.006122449
##
## $mids
## [1] 32.5 37.5 42.5 47.5 52.5 57.5 62.5
##
## $xname
## [1] "X[[i]]"
##
```

```
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $Scrambling
## $breaks
##  [1] 48 50 52 54 56 58 60 62 64 66 68
##
## $counts
##  [1]  3  6 18 37 43 47 30  6  5  1
##
## $density
##  [1] 0.007653061 0.015306122 0.045918367 0.094387755 0.109693878
##  [6] 0.119897959 0.076530612 0.015306122 0.012755102 0.002551020
##
## $mids
##  [1] 49 51 53 55 57 59 61 63 65 67
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $BounceBack
## $breaks
## [1] 12 14 16 18 20 22 24 26
##
## $counts
## [1]  2 20 36 53 44 26 15
##
## $density
## [1] 0.005102041 0.051020408 0.091836735 0.135204082 0.112244898 0.066326531
## [7] 0.038265306
##
## $mids
## [1] 13 15 17 19 21 23 25
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $PuttsPerRound
## $breaks
##  [1] 27.8 28.0 28.2 28.4 28.6 28.8 29.0 29.2 29.4 29.6 29.8 30.0 30.2
##
## $counts
##  [1]  1  1  3 11 19 32 34 38 23 14  9 11
##
## $density
##  [1] 0.02551020 0.02551020 0.07653061 0.28061224 0.48469388 0.81632653
##  [7] 0.86734694 0.96938776 0.58673469 0.35714286 0.22959184 0.28061224
##
## $mids
##  [1] 27.9 28.1 28.3 28.5 28.7 28.9 29.1 29.3 29.5 29.7 29.9 30.1
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

We can see from above plots that all of the predictors are normally distributed. Thus, we need not transform the variables.

Hence, I agree with the statistician's recommendation to apply log transform to the response variables only, but not the predictor variables.

b) Ans:

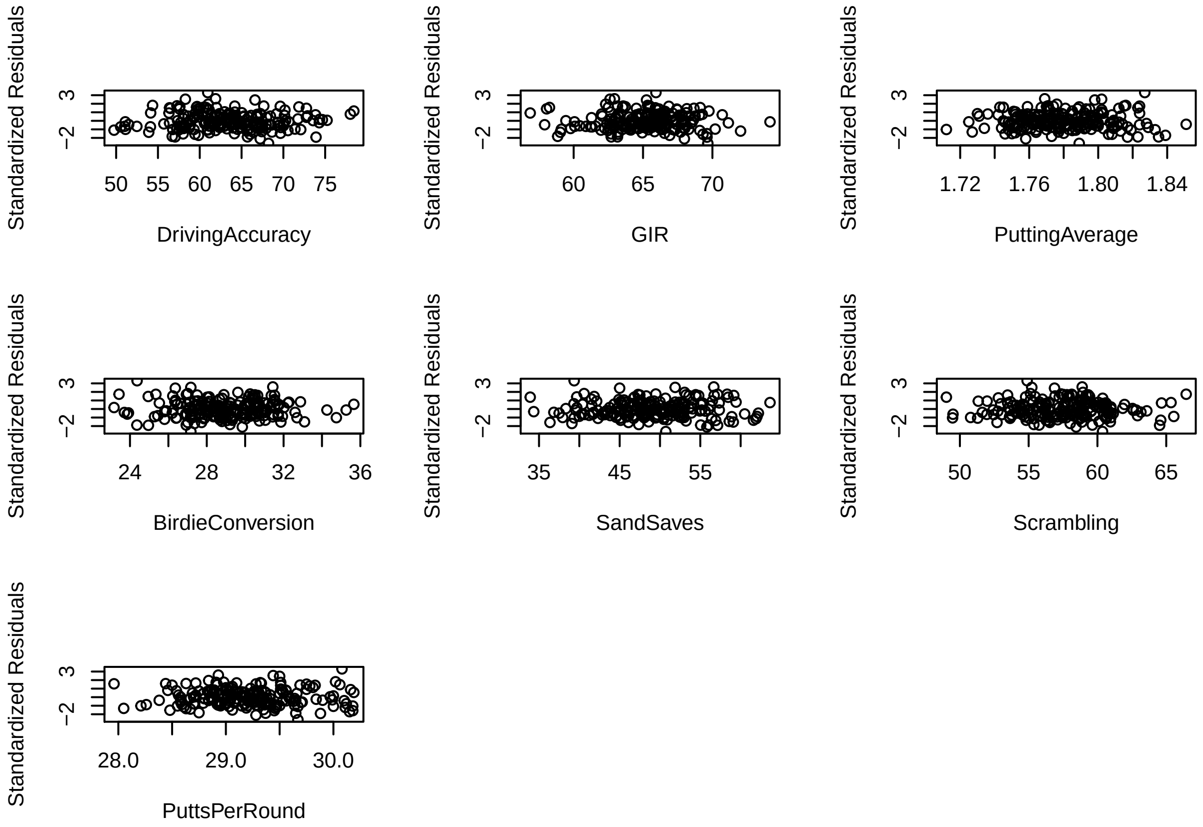
Let us now make a full regression model containing all seven potential predictor variables listed above.

```
mfull<-lm(formula=log(PrizeMoney)~DrivingAccuracy
+GIR
+PuttingAverage
+BirdieConversion
+SandSaves
+Scrambling
+PuttsPerRound
)
summary(mfull)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + SandSaves + Scrambling + PuttsPerRound)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71949 -0.48608 -0.09172  0.44561  2.14013
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.194300   7.777129   0.025 0.980095
## DrivingAccuracy  -0.003530   0.011773  -0.300 0.764636
## GIR               0.199311   0.043817   4.549 9.66e-06 ***
## PuttingAverage   -0.466304   6.905698  -0.068 0.946236
## BirdieConversion  0.157341   0.040378   3.897 0.000136 ***
## SandSaves         0.015174   0.009862   1.539 0.125551
## Scrambling        0.051514   0.031788   1.621 0.106788
## PuttsPerRound    -0.343131   0.473549  -0.725 0.469601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6639 on 188 degrees of freedom
## Multiple R-squared:  0.5577, Adjusted R-squared:  0.5412
## F-statistic: 33.87 on 7 and 188 DF,  p-value: < 2.2e-16
```

We can observe the coefficient values from the model above. We see that only GIR and BirdieConversion is significant variables in this case. Now, let us plot the standardized residuals for different variables.

```
StanRes1 <- rstandard(mfull)
par(mfrow=c(3,3))
plot(DrivingAccuracy,StanRes1, ylab="Standardized Residuals")
plot(GIR,StanRes1, ylab="Standardized Residuals")
plot(PuttingAverage,StanRes1, ylab="Standardized Residuals")
plot(BirdieConversion,StanRes1, ylab="Standardized Residuals")
plot(SandSaves,StanRes1, ylab="Standardized Residuals")
plot(Scrambling,StanRes1, ylab="Standardized Residuals")
plot(PuttsPerRound,StanRes1, ylab="Standardized Residuals")
```



We begin by looking at the condition below.

$$E(Y \mid X = x) = g(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) \qquad (6.6)$$

and

$$E(X_i \mid X_j) \approx \alpha_0 + \alpha_1 X_j \qquad (6.7)$$

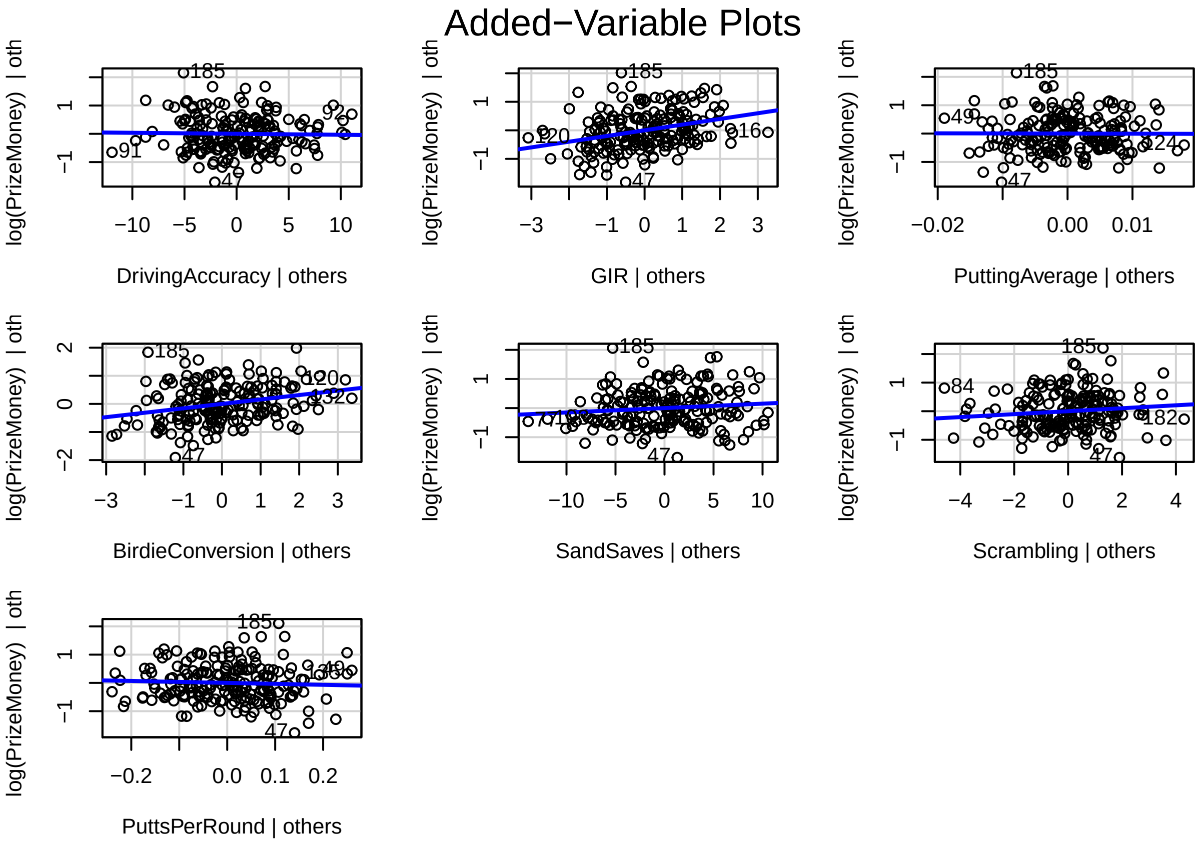
We begin by looking at the condition(6.7) above. We can see from the first scatterplot that the predictors seem to be related to each other linearly at least approximately.

Let us now observe added variable plots now.

```
library(MASS)
library(car)
```

```
## Loading required package: carData
```

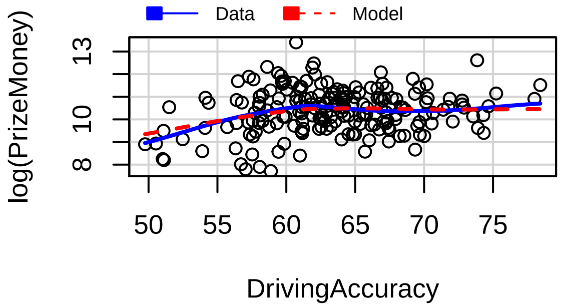
```
avPlots(mfull)
```

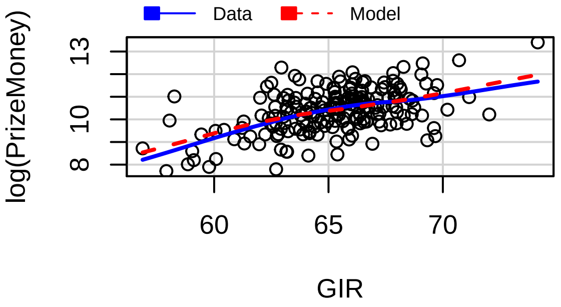


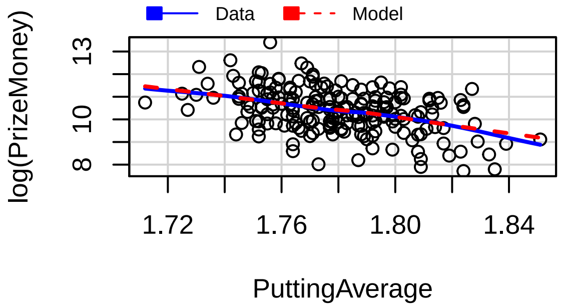
We can see from above plots that only two predictors (GIR and BirdieConversion) produce statistical significance and rest of the predictors don't show the same.

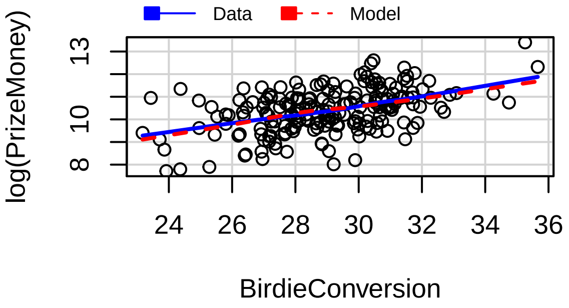
Let us now plot marginal plots.

```
par(mfrow=c(2,2))
mmp(mfull,DrivingAccuracy)
mmp(mfull,GIR)
mmp(mfull,PuttingAverage)
mmp(mfull,BirdieConversion)
```

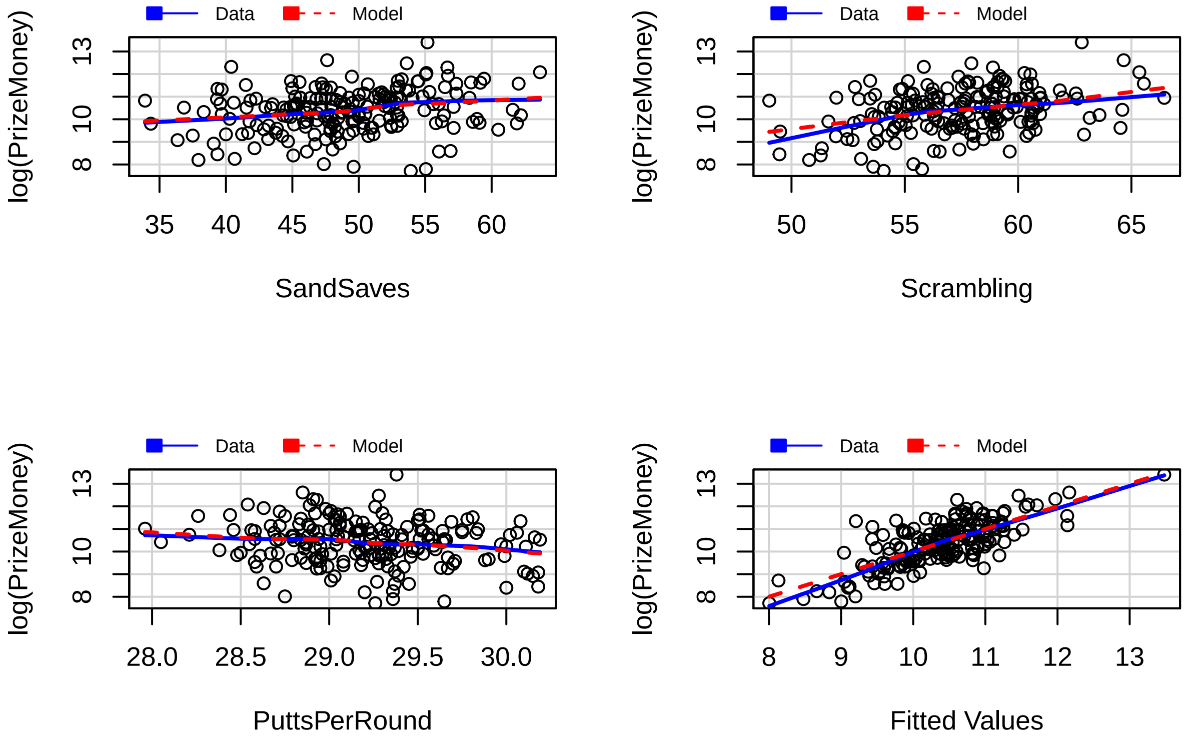




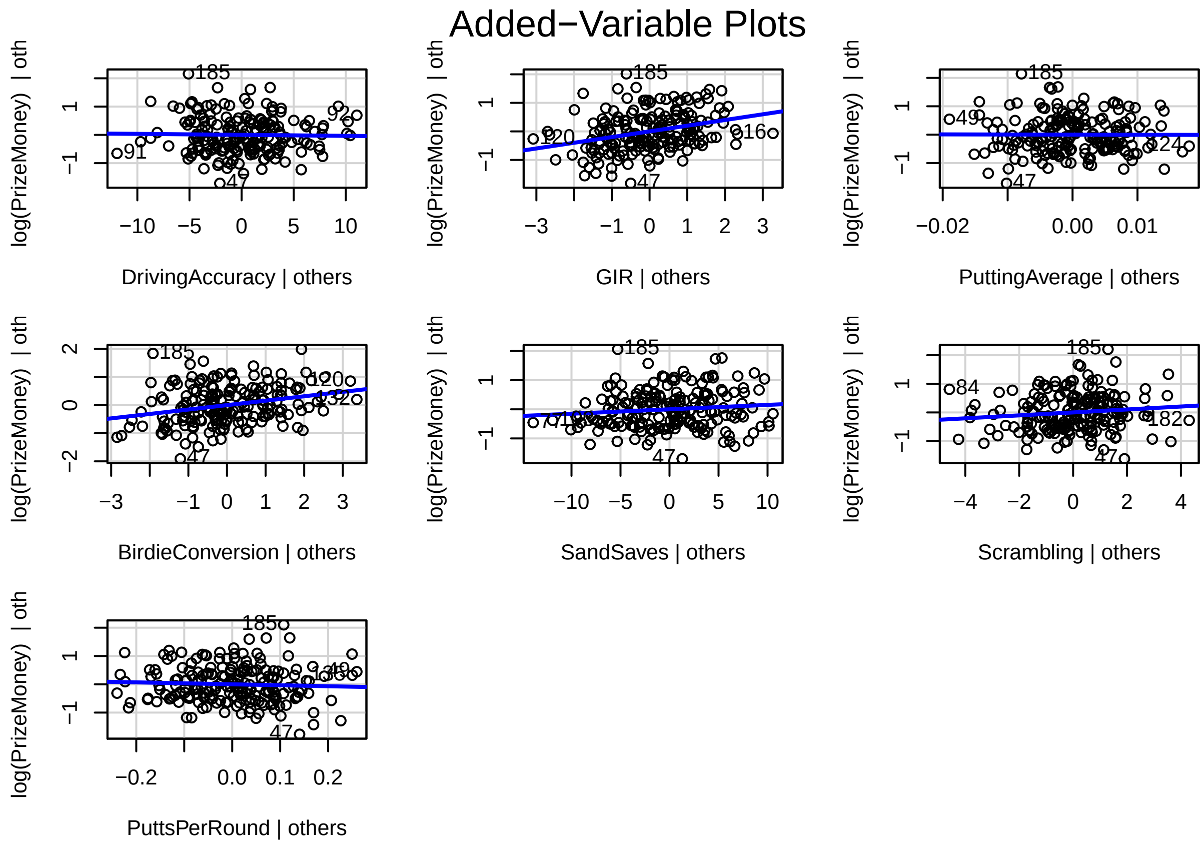




```
#par(mfrow=c(2,2))
mmp(mfull,SandSaves)
mmp(mfull,Scrambling)
mmp(mfull,PuttsPerRound)
mmp(mfull,mfull$fitted.values,xlab="Fitted Values")
```



```r
# Added-variable plots
avPlots(mfull)
```



All the lines are quite close to each other, so it seems the model is a good fit.

c) Ans:

We can declare all points are leverage points which is greater than 0.0816 (i.e., 2(p+1)/n = 16/196 = 0.0816). Therefore, the cases 40,70,77,91,120,168 and 178 appears to be highly influential in the model(i.e., they are considered to be high leverage points). All these points are away from average value of predictors and exercise greater effect on the model. These points should be seperately analysed.

d) Ans:

This is one of a valid model which can be improved very much. We haven't removed the insignificant predictors such as PuttingAverage. The model may also encounter type S and type m errors. Further investigation should be done in order to find out the proper relationship between the predictors and response. We also see that F-statistic(33.7) and R-squared(0.55) is also not favorable.

The term $1|1-R^2j$ is called the j^{th} variance inflation factor(VIF). Where ${R^2}_j$ denote the value of R^2 obtained from the regression of x_j on the other $x's$.

The variance inflation facotors for the valid model are:

```
## DrivingAccuracy              GIR   PuttingAverage BirdieConversion
##        1.796616         6.294969        12.900789         3.511898
##       SandSaves       Scrambling    PuttsPerRound
##        1.461506         4.470203        19.355667
```

A number of these variance inflation factors exceed 5, the cut-off often used, and so the associated regression coefficient are poorly estimated due to multicollinearity.

e) Ans:

In case of multiple regression, there might be some interplay between variables which can be seen in the scatterplot matrix. When this happens, the effect of variables is hard to be studied directly and it is difficult to evaluate the distinct effects of predictors on the response variable.

Thus, I would not recommend the approach of removing all predictors with insignificant t values because of the issues associated with the added variable plots, and multicollinearity. Whiel removing variables we might remove predictors which actually are good predictors. To remove any variable, we need to perform intensive analysis.

Project Milestone

Submitted with Dhurba Neupane