**honor statement: "I have completed this work independently.**

**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 a job in data science.**

When building a model by statistics or empirical rule to predict our target parameters, we may not build a perfect model to predict exactly accurate results , so all models are wrong.Depending on what is the accuracy that we ask for to decide whether the model is useful. According to my own working experience, in the apparel manufacturing industry, we have to forecast the sales volume three months later based on the sales data in the past few years, so that the company can plan in advance to produce apparel to be shipped to customers in three months. Company can not produce too many pieces in advance, otherwise it will cause bad debts, neither can not produce too few pieces in advance, otherwise the company will cause loss because of the lack of inventory to sell to its customers.
So we have to build a model to predict sales three months later, and decide whether the model works based on the loss that the company can accept..

**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.**

First thing I want to do to get a sense of the data is plot the dataset.
We can see a high correlation between PuttingAverage and Birdieversior, also between PuttingAverage and PuttsPerRound where I might leave concern of multicollinear in the dataset.
Then I would look at the correlation table to check if there is any value more than 90%. The correlation between PuttingAverage and PuttsPerRound is 0.79.

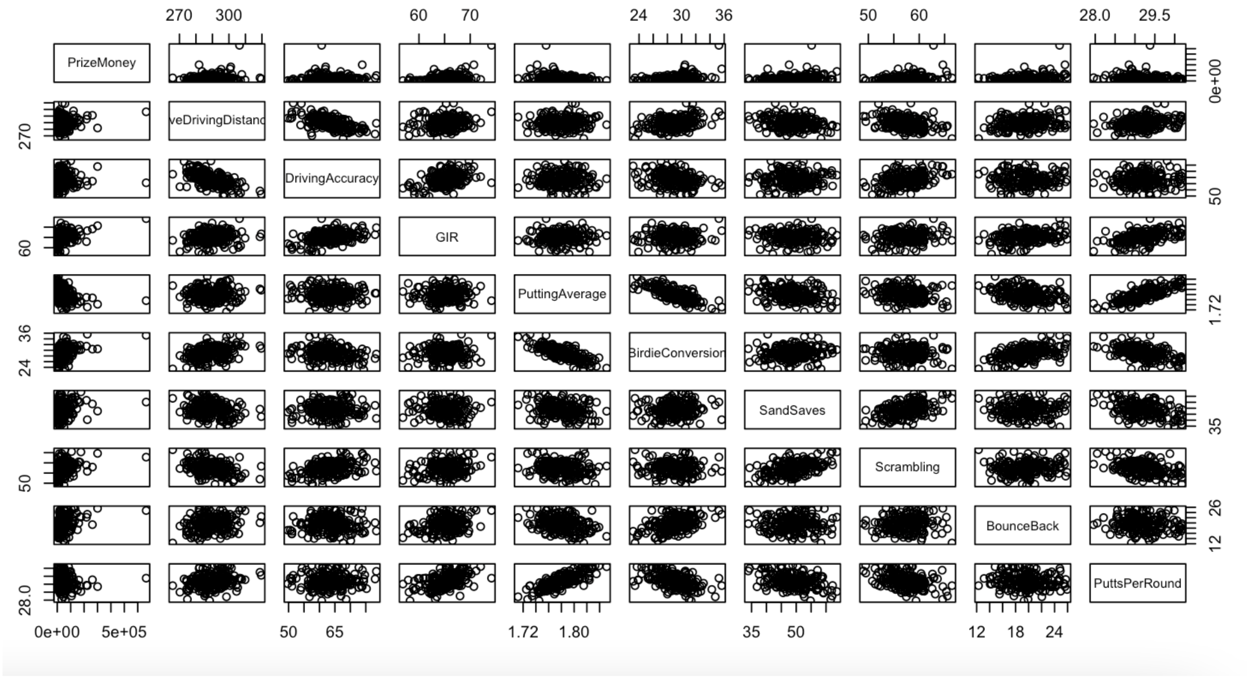
Next step, I would build up a model to see the F-test , R-square and T-test.
In this case, the F-test looks good to reject the null hypothesis and R-square is 0.54.
The P- value of the variables PuttingAverage and PuttsPerRound , we would think these two variables are not a good predictor.
However, when we check the plot, the correlation between PuttingAverage and logPrizeMoney is strong , so it might be because it is overwhelmed by other predictors, so it could be the impact of the multicollinear.

I would try to remove the variable of PuttingAverage first, the R-squared has been improved a little be to 0.5423 and if I remove  PuttsPerRound the r-squared is 0.5416 , so I will remove PuttingAverage and keep PuttsPerRound.

Next step, I would check VIF in the model, what we are worried about is any value more than 10. We would see the VIF value of PuttingAverage is  13.44 and VIF value of PuttsPerRound is 20.06 , it verifies there is multicollinear between these two variables, so I would remove PuttingAverage.After removing PuttingAverage, the VIF value of the rest of the variables look good.

Then I would remove the variable DrivingAccuracy that has high p-value, afterward I would remove bounceback.
To this step, whatever I try to remove the variables from the model , I can not improve r-squared, so my final model is > model<-lm(logPrizeMoney~GIR +BirdieConversion +SandSaves+Scrambling +PuttsPerRound ,data =  pgatour2006) and r-squared is 0.5459 .

```
pgatour2006$logPrizeMoney<-log(pgatour2006$PrizeMoney)
cor(pgatour2006)
```



```
> cor(pgatour2006)

> vif(model)
```

```
 DrivingAccuracy        GIR                  BirdieConversion       SandSaves       Scrambling
       1.673130       3.565504              3.054349               1.466007        2.755993
      BounceBack    PuttsPerRound
       1.462934       5.094736
```

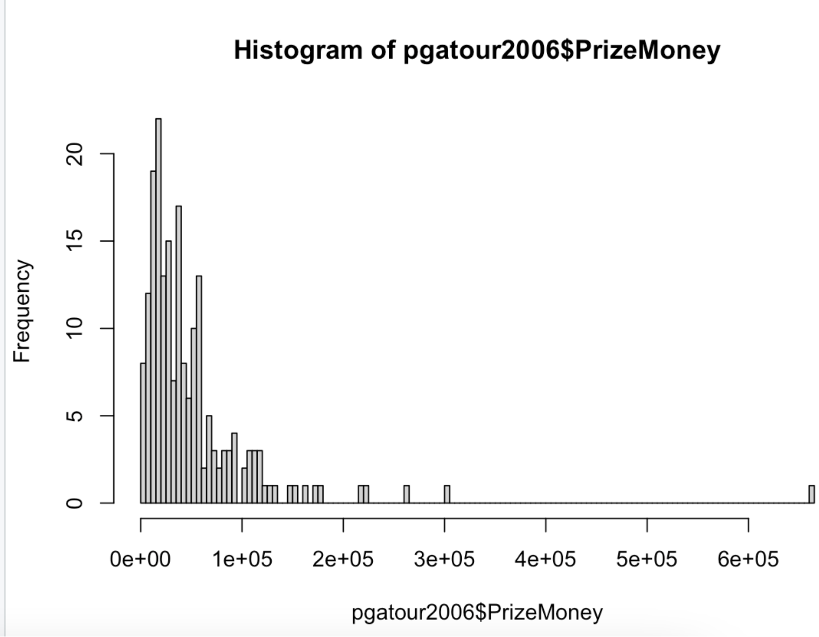
```
> model<-lm(logPrizeMoney~GIR +BirdieConversion +SandSaves+Scrambling +PuttsPerRound
,data =  pgatour2006)
> summary(model)

Call:
lm(formula = logPrizeMoney ~ GIR + BirdieConversion + SandSaves +
    Scrambling + PuttsPerRound, data = pgatour2006)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71291 -0.48168 -0.09097  0.44843  2.15763

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.583181   7.158721  -0.081   0.9352
GIR               0.197022   0.028711   6.862 9.31e-11 ***
BirdieConversion  0.162752   0.032672   4.981 1.41e-06 ***
SandSaves         0.015524   0.009743   1.593   0.1127
Scrambling        0.049635   0.024738   2.006   0.0462 *
PuttsPerRound    -0.349738   0.230995  -1.514   0.1317
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6606 on 190 degrees of freedom
Multiple R-squared:  0.5575,  Adjusted R-squared:  0.5459
F-statistic: 47.88 on 5 and 190 DF,  p-value: < 2.2e-16
```

**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?**

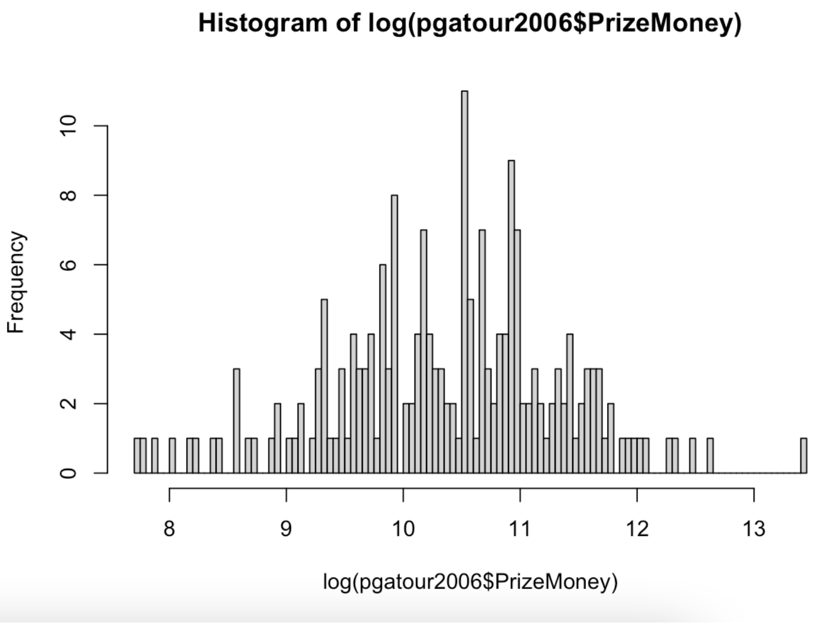
Before the log transformation, we could not detect the normality of the data on histogram graph, and its adjusted R-square is 0.381.

After the log transformation, the graph pulled in some data on the right and the data on the right which could be the outlier is less significant in the log graph than in the regular graph. It performs as a normal distribution on histogram graph and its adjusted R-square is 0.54 which is higher than before log transformation.

hist(pgatour2006$PrizeMoney, breaks = 100)



hist(log(pgatour2006$PrizeMoney), breaks = 100)



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

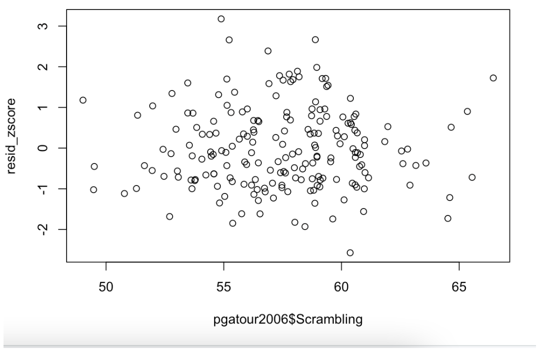
```
> sum(model $residuals)
[1] 4.440892e-15
```

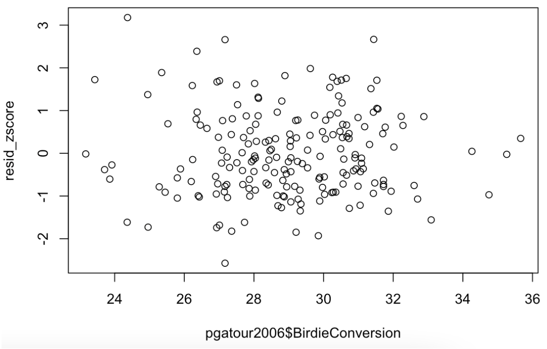
Check the sum of residual which is close to zero sum 4.440892e-15, though our assumption should be the sum of residual which is equal to zero.

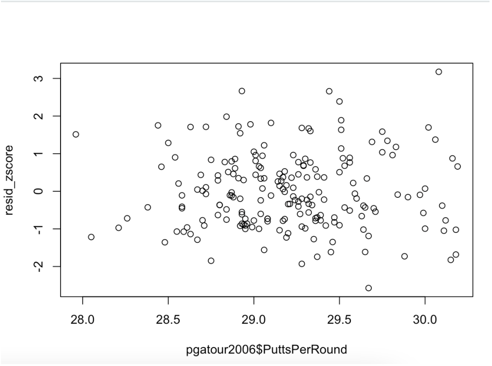
Next is to check the histogram of the  residuals. I see a normal curve, but there are a few outliers that are further away than we might want.
Then I ‘ll make a z-score normalization of residuals and I’ll get 95% of the residuals that are within 2 standard deviations.
I would check the independence by  Durbin Watson Test, p -value of Durbin Watson value is 0.244 which means the residuals are dependent on each other.

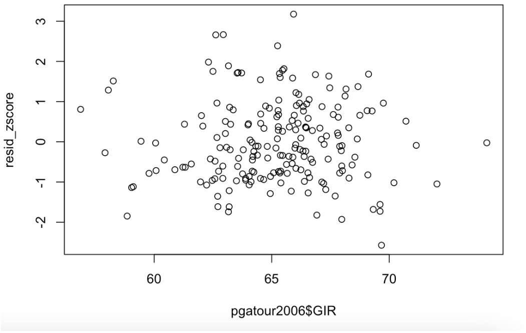
```
>plot(pgatour2006$AveDrivingDistance,resid_zscore)
> plot(pgatour2006$BirdieConversion,resid_zscore)
> plot(pgatour2006$Scrambling,resid_zscore)
> plot(pgatour2006$PuttsPerRound,resid_zscore)
> plot(pgatour2006$GIR,resid_zscore)
```

There is no huge difference from the left side all the way to the right side.
They look like healthy graphs, 95% of the residuals are within two standard deviations. They should be homoscedastic.

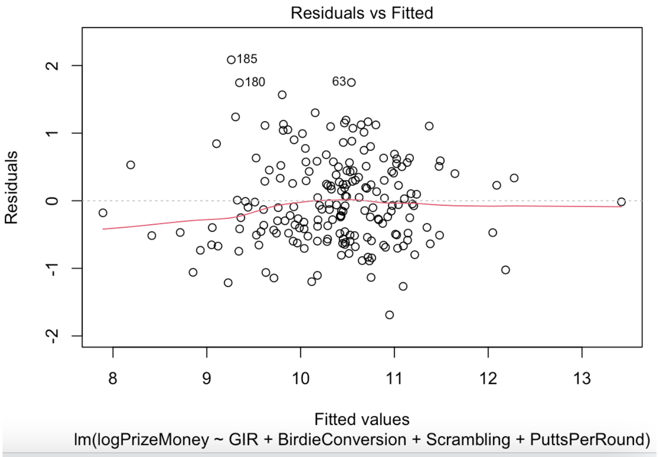




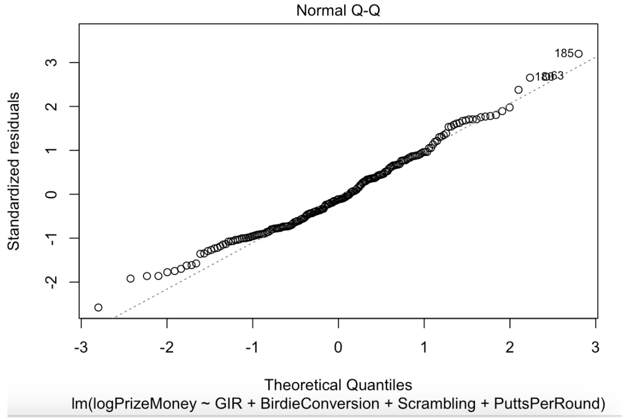




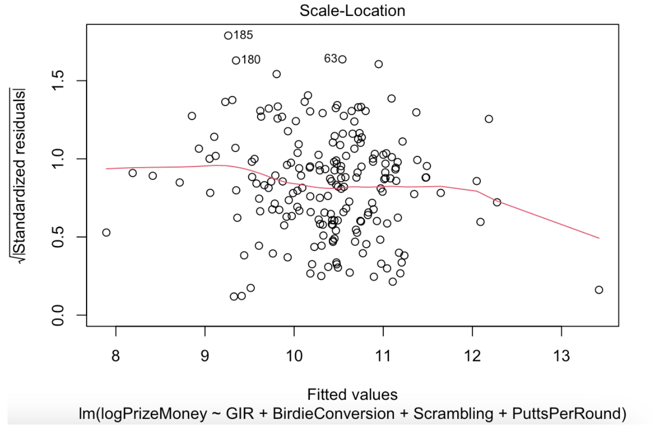
Also plot the model
There is no obvious trend on the plot of residuals vs fitted, it looks healthy.



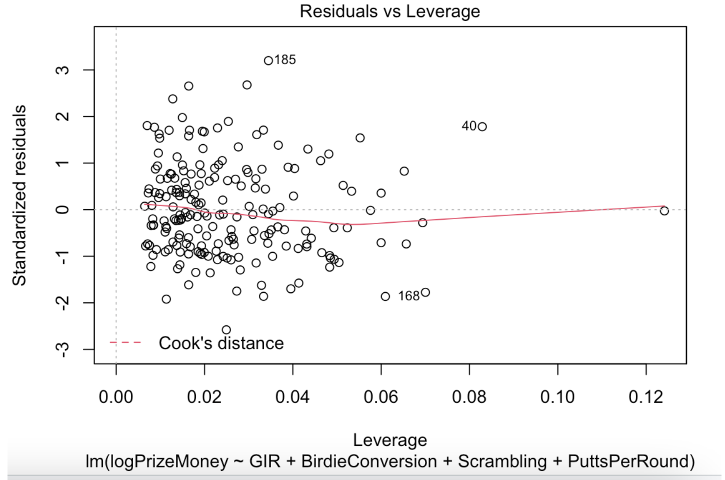
The Q-Q plot looks normal, the dots are lined on the diagonal line.



The plot of  standardized residuals looks healthy.



It shows plot leverage of regression line

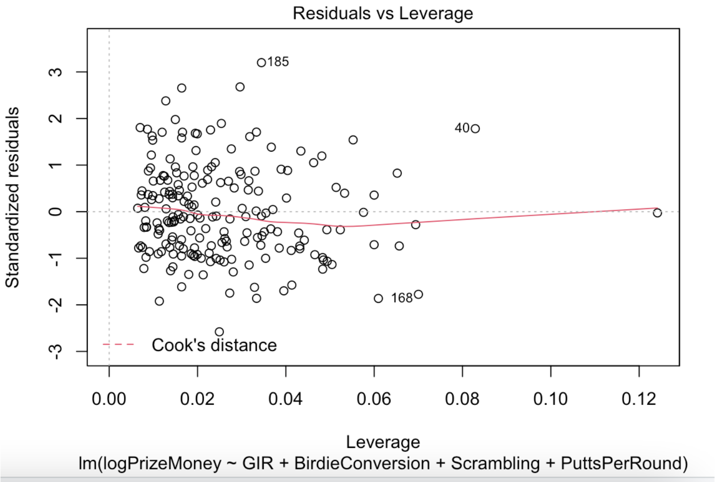


**d. (10 points) Analyze if there are any outliers and/or influential points. If there are points in the dataset that need to be investigated, give one or more reasons to support each point chosen. Discuss your answer.**

Here is the standardized residuals  versus the Leverage, we can see there is an observation point that lies closest to the border of Cook's distance, but it doesn't fall outside of the dashed line, so I would say there are not any influential points in our regression model.

[1]Residuals vs Leverage plot definition:
Leverage refers to the extent to which the coefficients in the regression model would change if a particular observation was removed from the dataset.
Observations with high leverage have a strong influence on the coefficients in the regression model. If we remove these observations, the coefficients of the model would change noticeably.

If any point in this plot falls outside of Cook's distance (the red dashed lines) then it is considered to be an influential observation.



[1] Residuals vs Leverage plot definition:https://www.statology.org/residuals-vs-leverage-plot/