**DSC 423: Data Analytics and Regression**

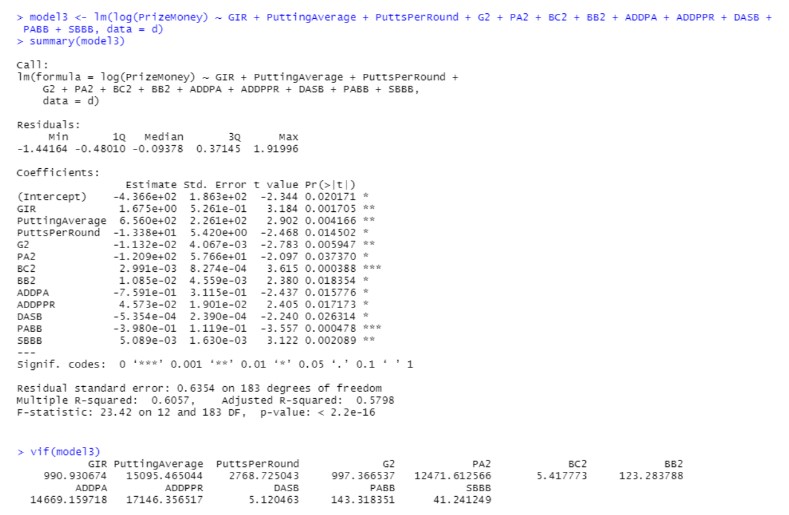
**Assignment 08**

**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.**
   1. **(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.**



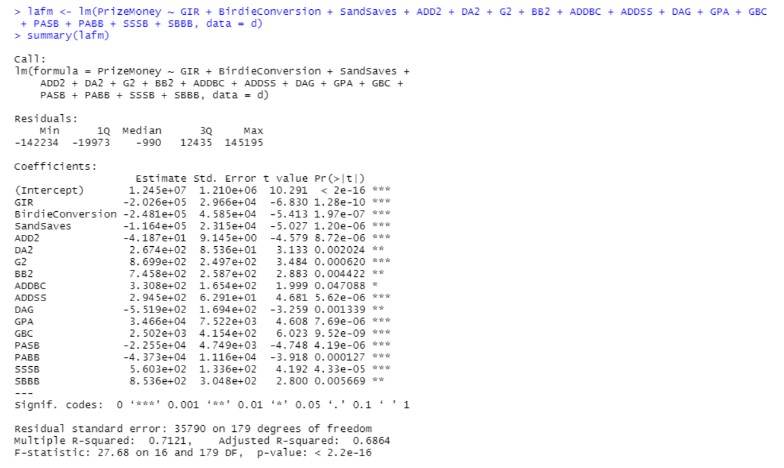
d$G2 <- d$GIR^2 d$PA2 <- d$PuttingAverage^2 d$BC2 <- d$BirdieConversion^2 d$PABB <- d$PuttingAverage \* d$BounceBack d$BB2 <- d$BounceBack^2 d$SBBB <- d$Scrambling \* d$BounceBack d$ADDPA <- d$AveDrivingDistance \* d$PuttingAverage d$ADDPPR <- d$AveDrivingDistance \* d$PuttsPerRound

* 1. **(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?**

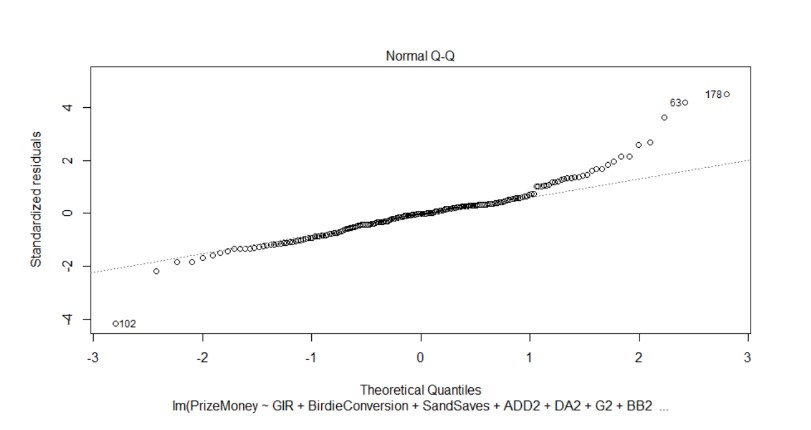
Last Assignment Final Model:

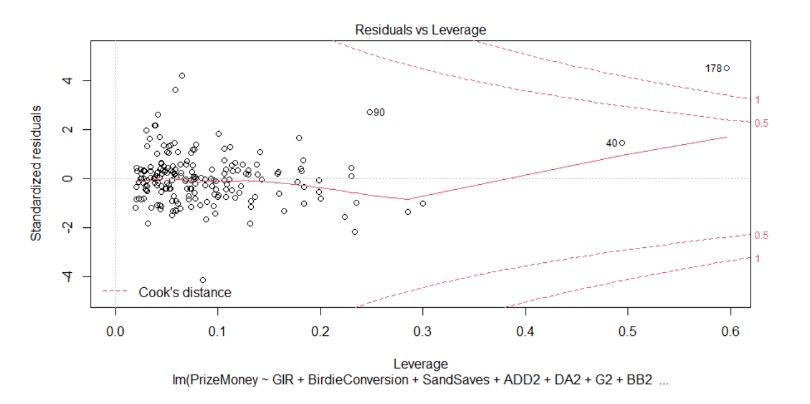
lafm <- lm(PrizeMoney ~ GIR + BirdieConversion + SandSaves + ADD2 + DA2 + G2 + BB2 +

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

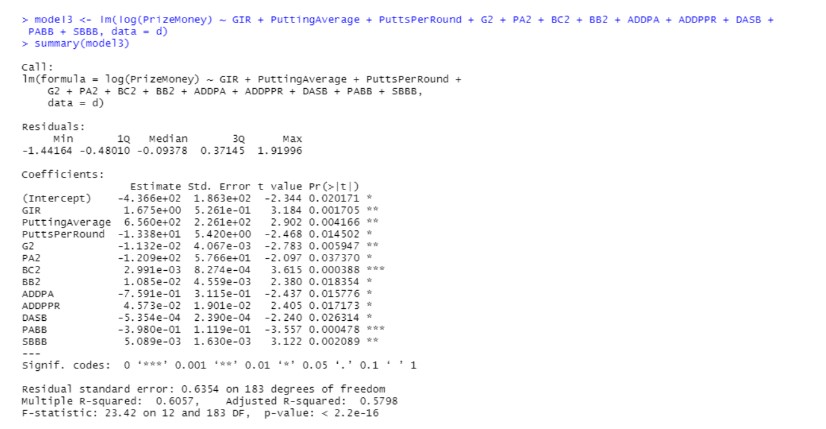


d$ADD2 <- d$AveDrivingDistance^2 d$DA2 <- d$DrivingAccuracy^2 d$G2 <- d$GIR^2 d$BB2 <- d$BounceBack^2 d$ADDBC <- d$AveDrivingDistance \* d$BirdieConversion d$ADDSS <- d$AveDrivingDistance \* d$SandSaves d$DAG <- d$DrivingAccuracy \* d$GIR d$GPA <- d$GIR \* d$PuttingAverage d$GBC <- d$GIR \* d$BirdieConversion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

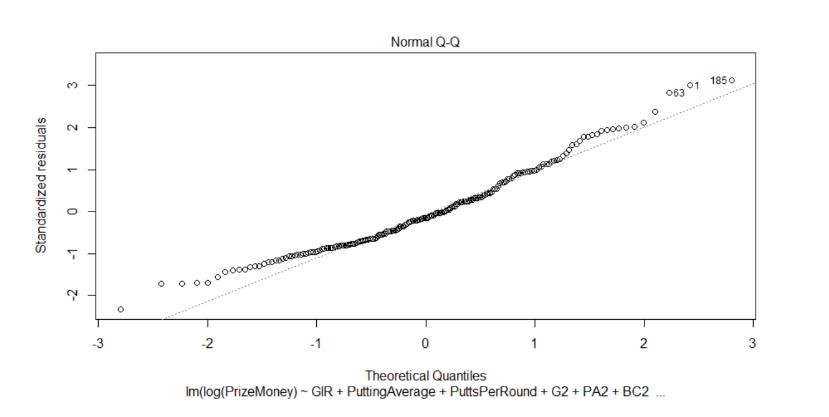


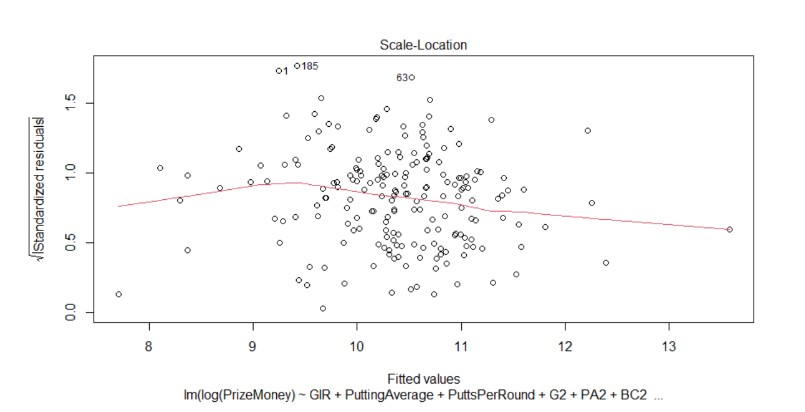


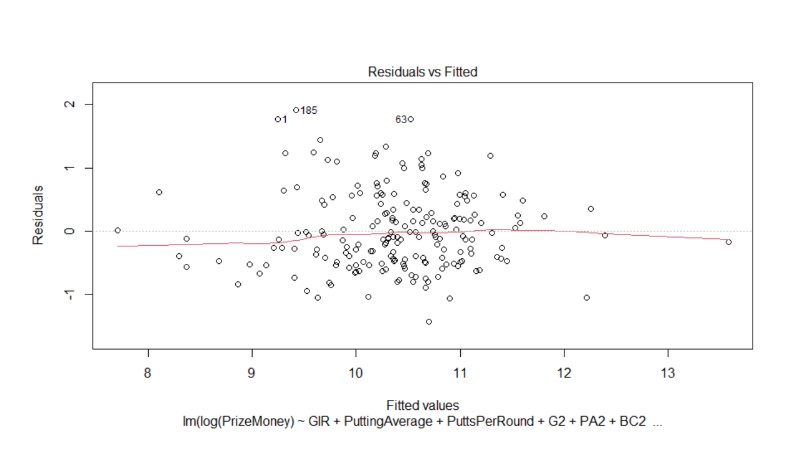
Present assignment Model:

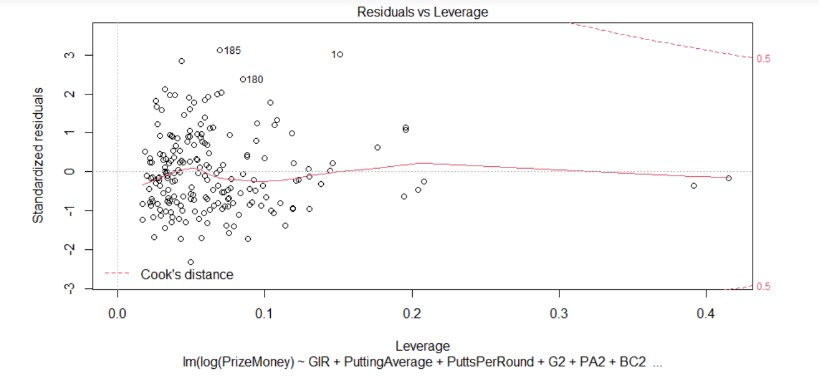


We can use summary(model) and plot to compare the two models (model). As can be observed from the comparison, the adj-R2 value of the prior model without the log transformation is higher. On the other hand, the present model log transforms the normalQQ data. The graph looks like a direct diagonal In 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.



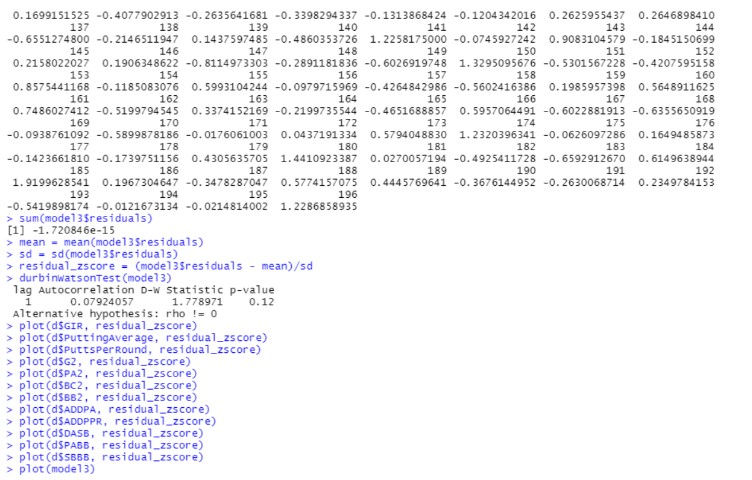
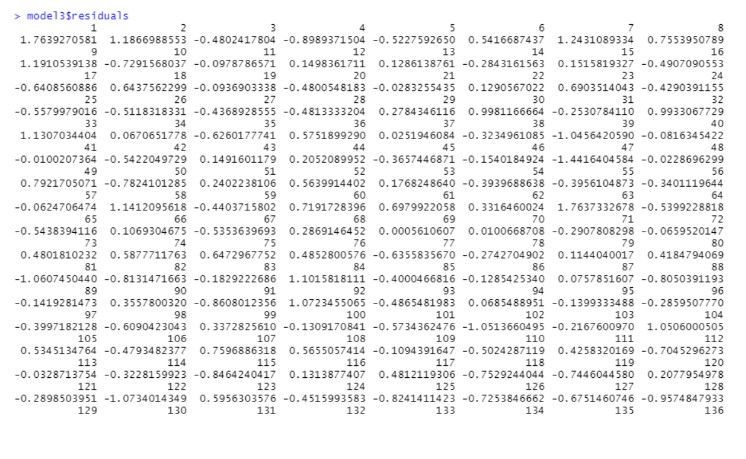


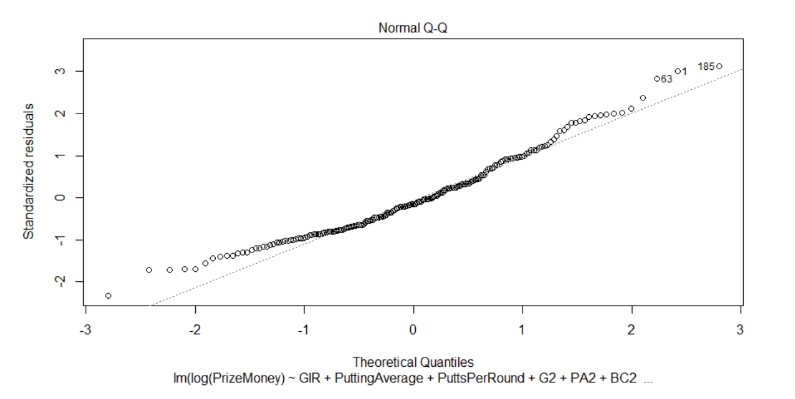


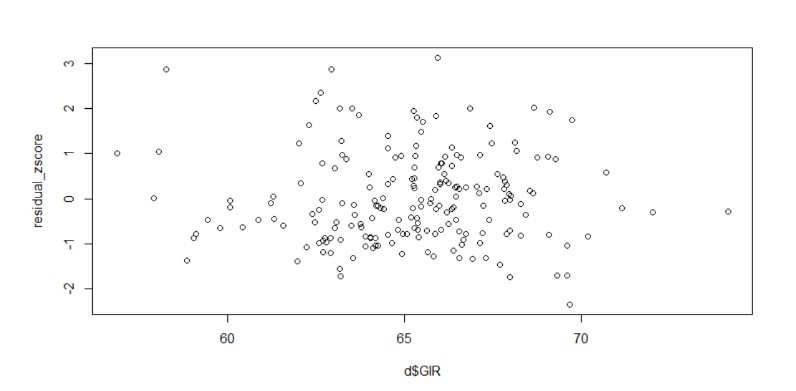


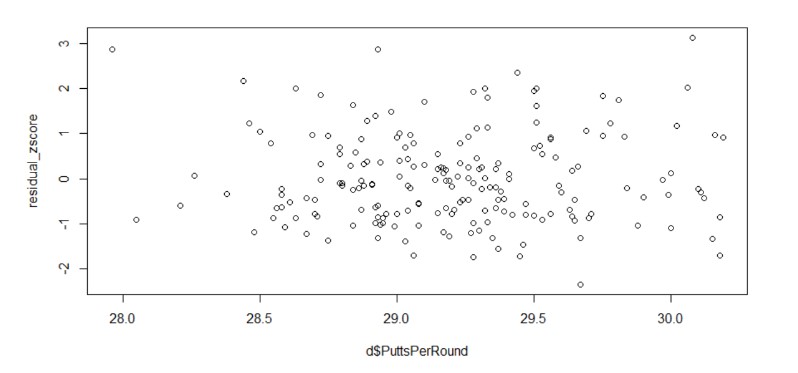
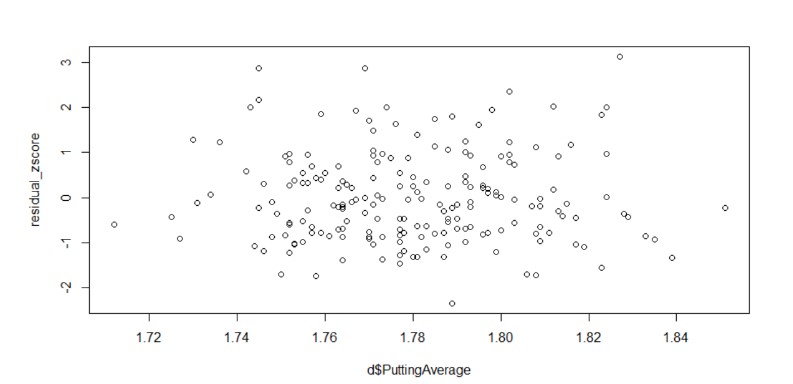
1. **(10 points) Analyze and discuss the residual plots.**

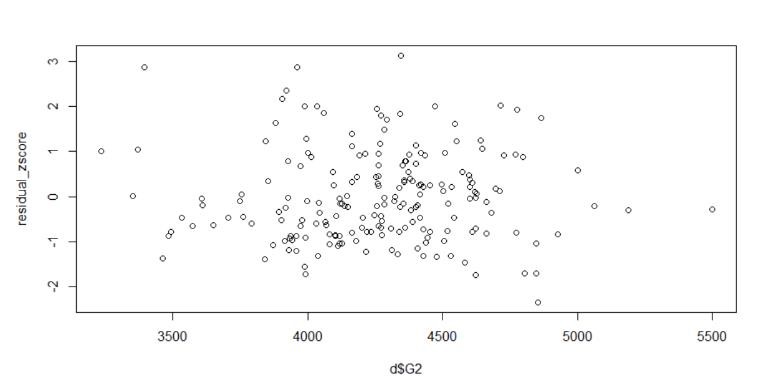
In our current log model of studentized residual graph, the model is normally distributed along a linear line. The zero line has x variable graphs scattered about it at random. They stand for independence and continual variance. There might be outliers in the graphs of the x variables because the studentized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

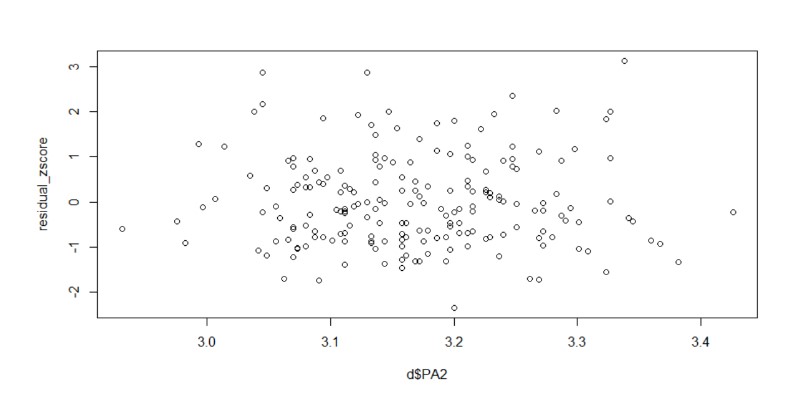


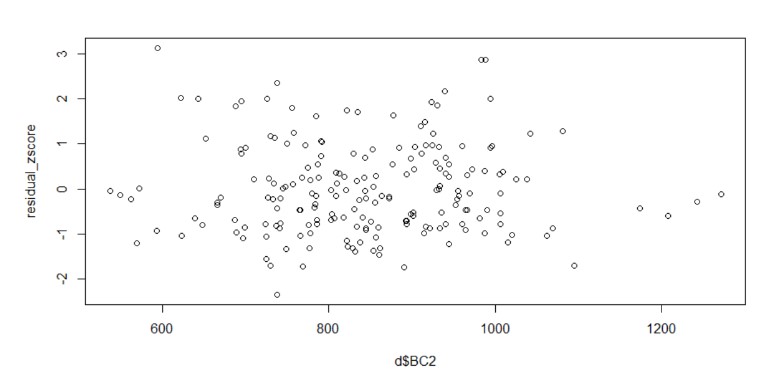


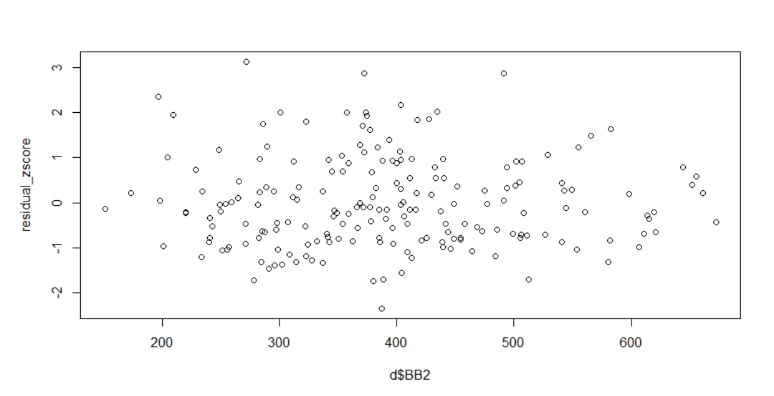


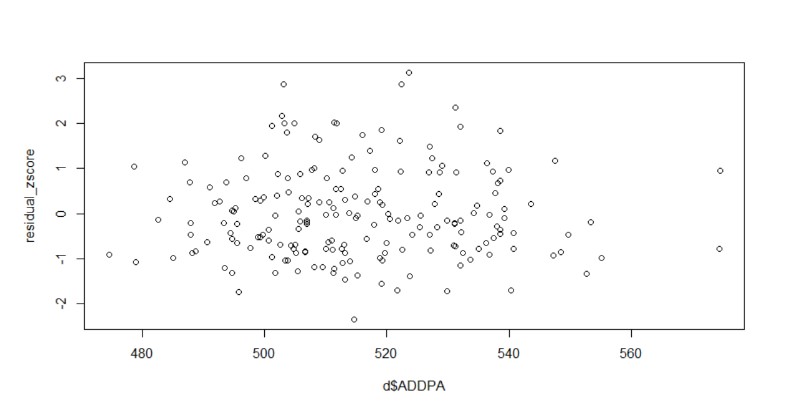


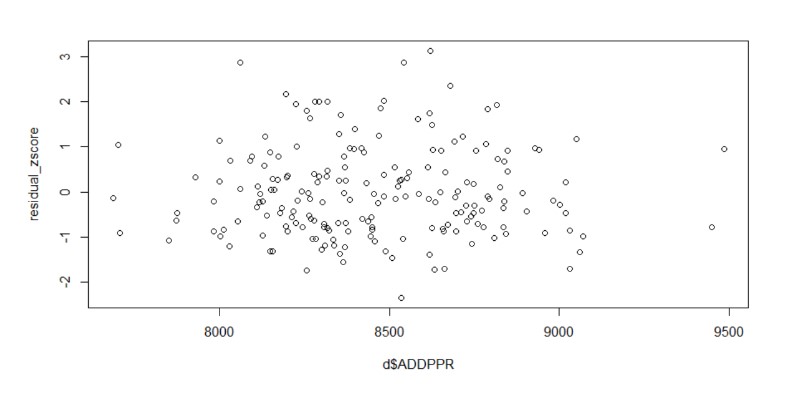


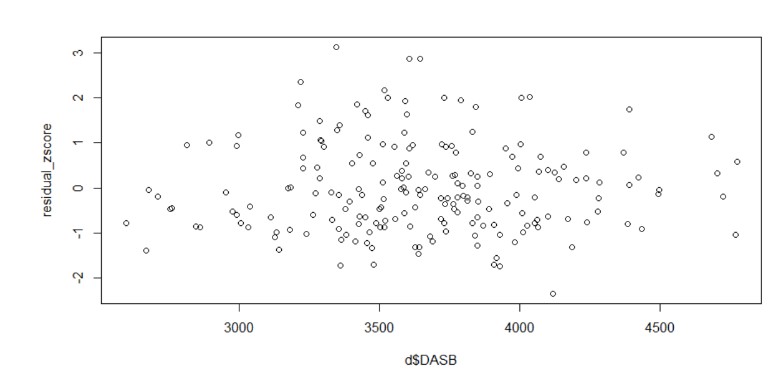


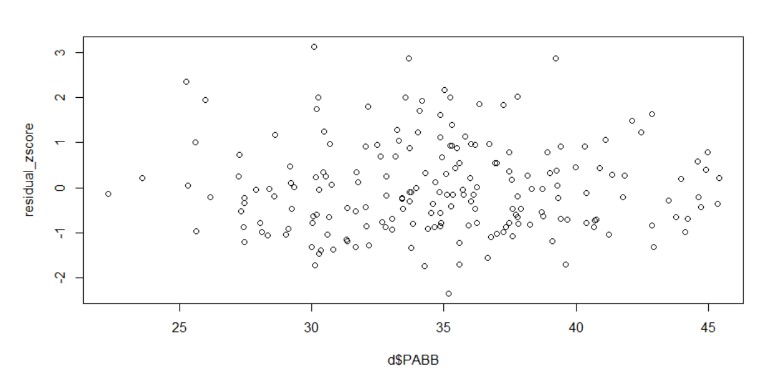


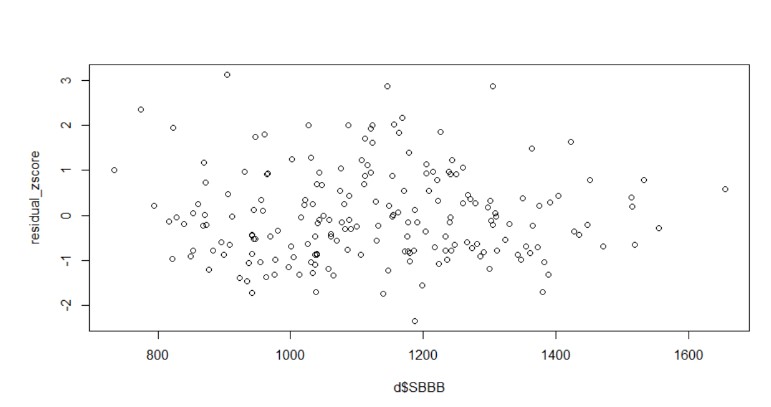












1. **(10 points) Analyze if there are any outliers and/or influential points. If there are points in the dataset that need to be investigated, give one or more reason to support each point chosen. Discuss your answer.**

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.