Title: Assignment 2

Travis Ho

2023-10-14

Setup

```
#Load libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Data

```
#Load NBA stats dataset
nba <- read.csv("./data/nba_2022.csv")
dim(nba)
```

```
## [1] 467  52
```

This dataset was obtained from the Kaggle dataset called "NBA Player Salaries (2022-23 Season)" This is a dataset containing offensive and defensive statistics from all NBA players who played minutes in the 2022 season. The dataset consists of 467 players and has 52 different variables that include basic statistics such as points scored but also advanced statistics such as Value Over Replacement Player (VORP).

Research Question

In basketball, the assumption is that the closer you get to the basket, the easier it must be to make a shot. From this assumption, many people believe that because a two point shot is closer to the basket than a three point shot, people who shoot more two point attempts should make a higher percent of their shots. The question I am proposing is: In the 2022 season of the NBA, was a player's field goal percentage (percent make) related to two point rate, three point attempts, and free throw rate?

- H0: The field goal percentage is not related to any of the variables.
- HA: The field goal percentage is related to at least one of the variables.

Variables of Interest

The four major variables of interest in this analysis will be two point rate and field goal percentage. Two point rate is the percent of shots a player takes that are worth two points. This is a continuous variable ranging from 0.0 to 1.0. This is calculated from the variables 2 point attempts and 3 point attempts. It is 2 point attempts divided by the total attempts (addition of 2 point attempts and three point attempts).Three point attempts is the amount of three pointers a player takes per game. This is a continuous variable that is all positive rational numbers. Free throw rate is the percent of times a player takes free throws. This is a continous variable that ranges from 0.0 to 1.0. Field goal percentage is the percent of times a player makes a shot. This is a continuous variable ranging from 0.0 to 1.0.

Other variables that will be investigated are player name which is a categorical variable that is simply the player name. Another variable that will be highlighted is minutes played. This is a continuous variable that is any positive rational number on the range of 0 to 48. Games played is another variable. This is the amount of games a player played in the season. This is a continuous integer that ranges from 0 to 82.

Data Wrangling

```
#Create a new dataframe with only the variables we need
stats <- nba %>%
         select(Player.Name,
         MP,
         GP,
         X2PA,
         X3PA,
         FTr,
         FG.)
```

```
#Create two point rate variable
stats$twoPR <- stats$X2PA / (stats$X2PA + stats$X3PA)
```

```
#Filter out all players that played less than 10 minutes per game and played less than 20 games.
stats <- stats %>%
         filter(MP > 20.0 & GP > 21)
```

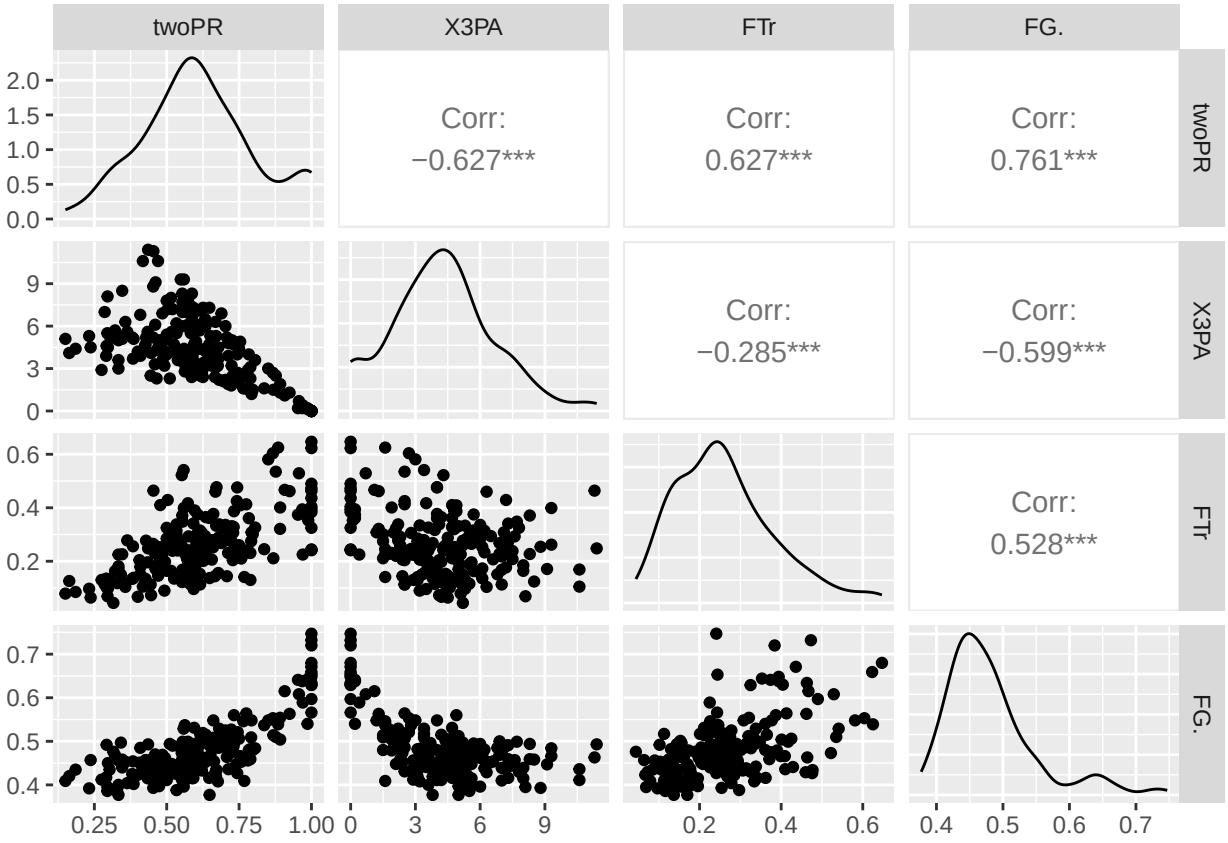
```
#Shorten the dataframe to only the variables we need
stats_reg <- stats %>%
  select(twoPR,
```

```
          X3PA,
          FTr,
          FG.)
```

```
#Transform the data
library(bestNormalize)
set.seed(1234)
selected_columns <- c("twoPR","X3PA","FTr","FG.")
stats_normal <- lapply(
  stats_reg[selected_columns],
  function(x){
    bestNormalize(x)$x.t
  }
)
#Convert it to a dataframe
stats_normal <- as.data.frame(stats_normal)
```

Analysis and Data Visualization (No Transformations)

```
#Lets take a look at the non transformed distributions and correlations
ggpairs(stats_reg)
```



None seem to have that strong of a correlation with field goal percentage. The highest is two point rate at 0.761.

```r
#Linear Model without Transformations
lm_nba <- lm(formula = FG. ~ ., data = stats_reg)
summary(lm_nba)
```

```
##
## Call:
## lm(formula = FG. ~ ., data = stats_reg)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.109691 -0.026325 -0.002264  0.022922  0.167118
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.373100   0.017447  21.385  < 2e-16 ***
## twoPR        0.190051   0.024541   7.744 4.56e-13 ***
## X3PA        -0.006362   0.001638  -3.884 0.000139 ***
## FTr          0.069420   0.032060   2.165 0.031534 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04272 on 202 degrees of freedom
## Multiple R-squared:  0.6124, Adjusted R-squared:  0.6066
## F-statistic: 106.4 on 3 and 202 DF,  p-value: < 2.2e-16
```

Adjusted R-squared of 0.6066 suggests that 60.66% of the variance of field goal percentage can be explained by two point rate, three point attempts, and free throw rate.

Assumptions

```r
#Import car library and check VIF
library(car)
```

Multicollinearity

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
vif(lm_nba)
```

```
##    twoPR     X3PA      FTr
## 2.574128 1.702530 1.699480
```

None of these VIF are particularly large suggesting that the multicollinearity is not a problem.

```
#Look at the mean and standard deviation of the residuals without transformation.
mean_resid <- mean(residuals(lm_nba))
sd_resid <- sd(residuals(lm_nba))
mean_resid
```

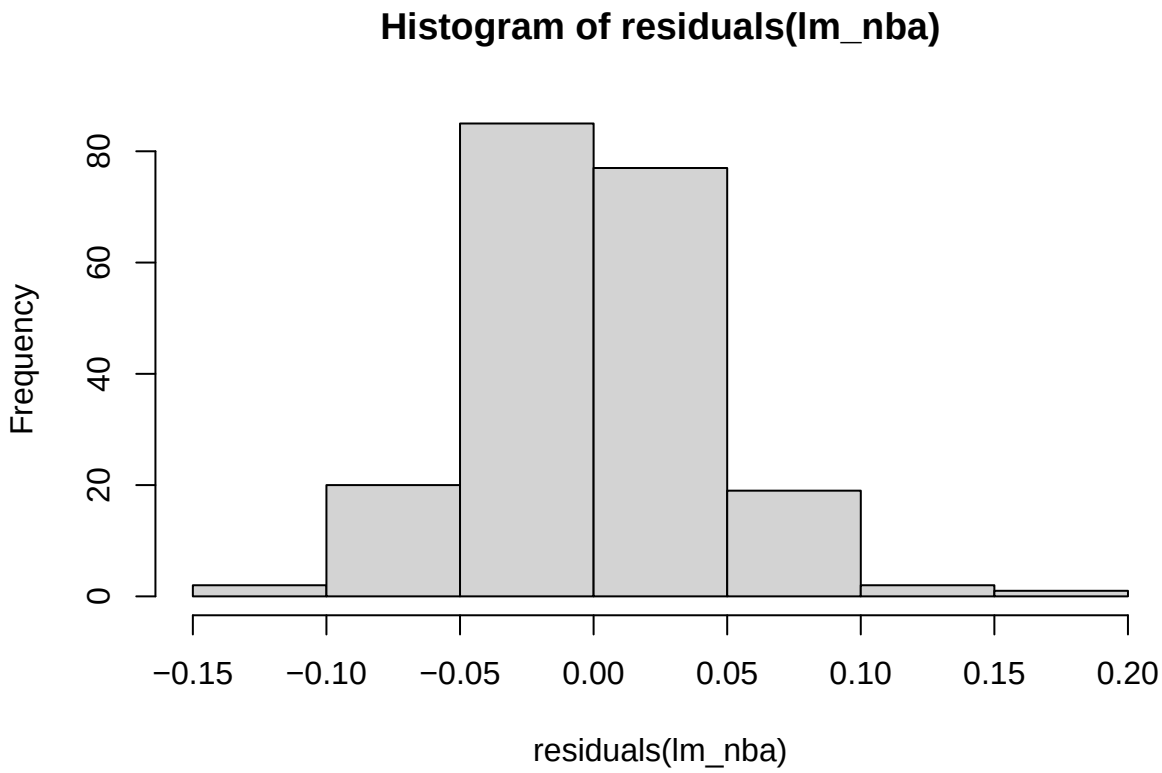
Residuals

```
## [1] -9.067457e-19
```

```
sd_resid
```

```
## [1] 0.04240396
```

```
#Get the histogram of the residuals
hist(residuals(lm_nba))
```



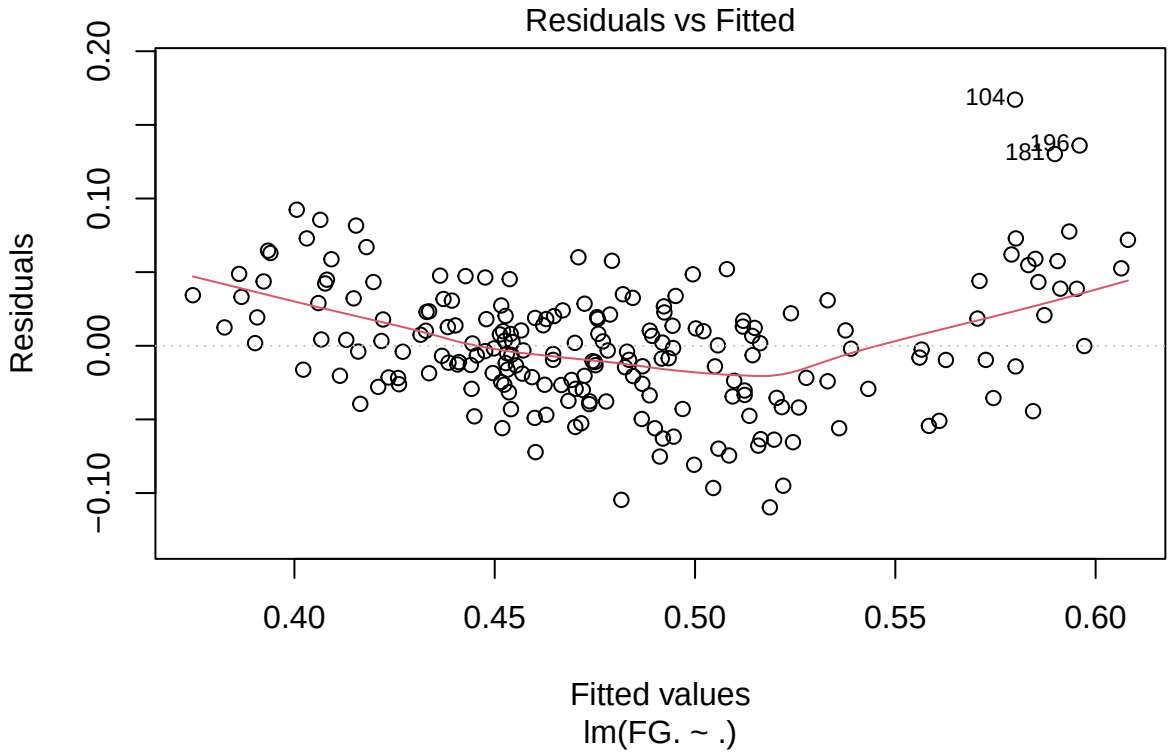
The residuals actually look somewhat normal but need to check with the Shapiro-Wilks Test.

```r
#Shapiro-Wilks Test
shapiro.test(residuals(lm_nba))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm_nba)
## W = 0.98447, p-value = 0.02298
```

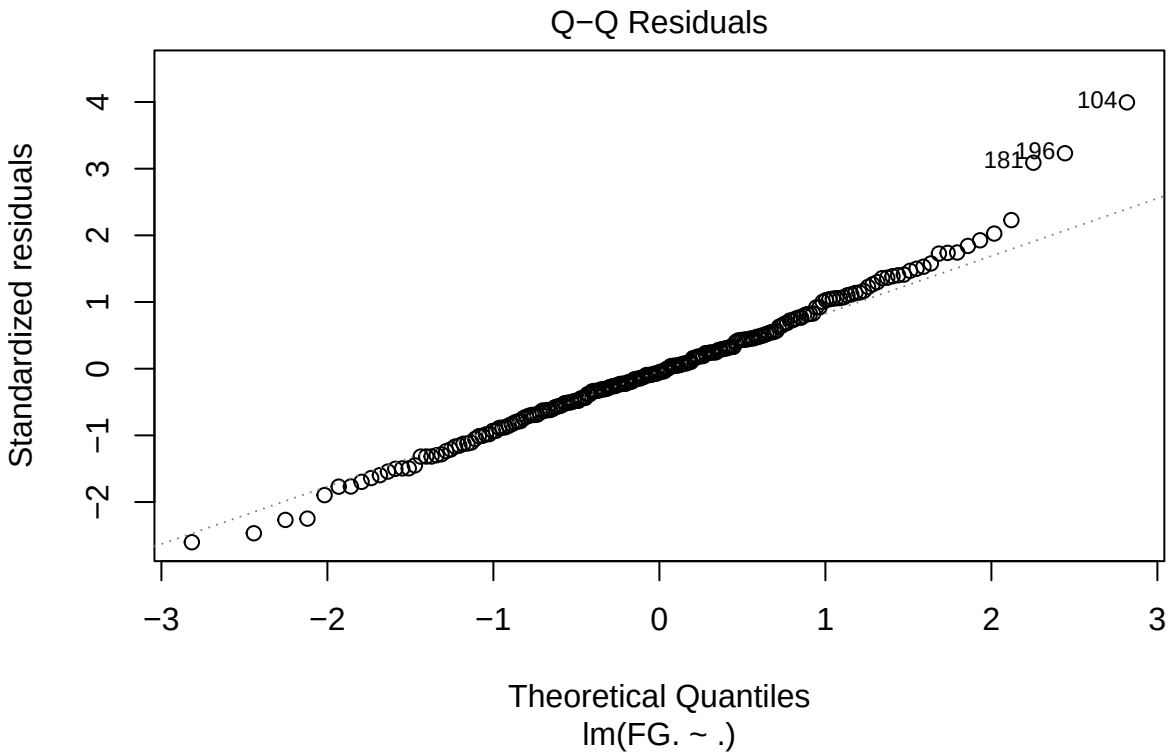
The p-value of 0.02298 is less than alpha 0.05 so we can reject the null hypothesis and conclude the residuals are not normally distributed.

```r
#Homoskedasticity
#Plot Fitted values on residuals
plot(lm_nba, which=1)
```



The medium to low end is very clumped around 0 while the high end has lots a variability. Does not look that great.

```r
#QQplot
plot(lm_nba, which=2)
```



The qq plot indicates that the values are normally distributed for the most part except on the high end where there is lots of variability. We will not try analysis of the data with transformation and see if the results are different.

Analysis and Data Visualization (With Transformation)

```
#Linear Model with Transformations
lm_nba_norm <- lm(formula = FG. ~ ., data = stats_normal)
summary(lm_nba_norm)
```

```
##
## Call:
## lm(formula = FG. ~ ., data = stats_normal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.04657 -0.40446  0.02228  0.44822  1.63847
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.681e-16  4.618e-02   0.000   1.0000
## twoPR        4.842e-01  7.295e-02   6.637 2.88e-10 ***
## X3PA        -2.489e-01  5.779e-02  -4.306 2.59e-05 ***
## FTr          1.439e-01  6.162e-02   2.335   0.0205 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6628 on 202 degrees of freedom
## Multiple R-squared:  0.5672, Adjusted R-squared:  0.5607
## F-statistic: 88.23 on 3 and 202 DF,  p-value: < 2.2e-16
```

Assumptions

```
#VIF of normalized data
vif(lm_nba_norm)
```

Multicollinearity

```
##    twoPR     X3PA      FTr
## 2.483416 1.558527 1.771790
```

```
#Mean and standard deviation of residuals of transformed data
mean_resid_norm <- mean(residuals(lm_nba_norm))
sd_resid_norm <- sd(residuals(lm_nba_norm))
mean_resid_norm
```

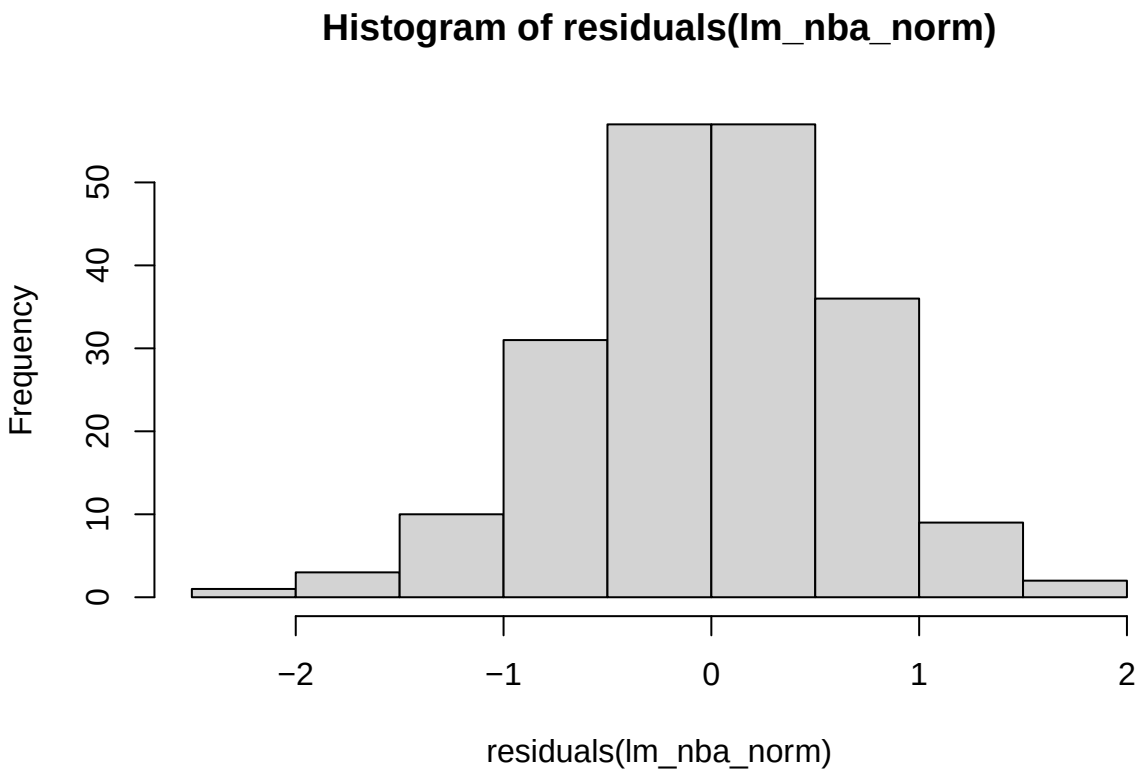
Residuals

```
## [1] 8.39381e-17
```

```
sd_resid_norm
```

```
## [1] 0.6579015
```

```
#Histogram of residuals of normalized data
hist(residuals(lm_nba_norm))
```



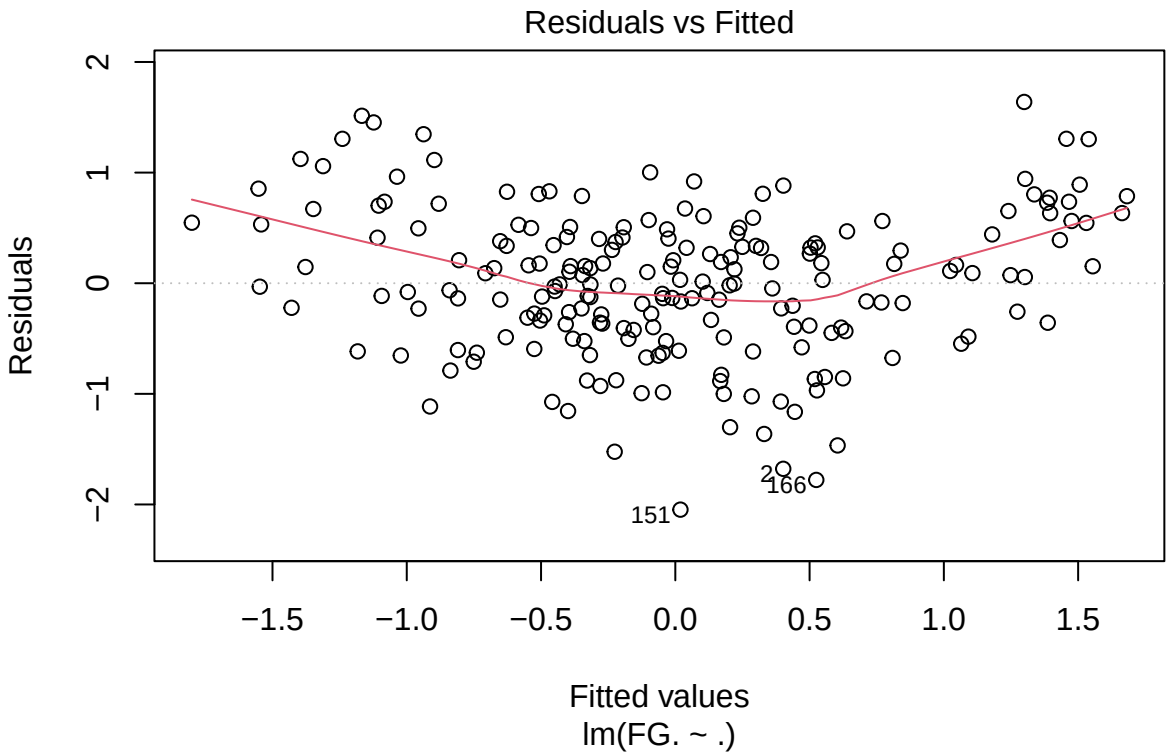
Looks similar to non-transformed data, might be a little better in terms of normality

```
#Shapiro-Wilks Test
shapiro.test(residuals(lm_nba_norm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm_nba_norm)
## W = 0.99525, p-value = 0.769
```

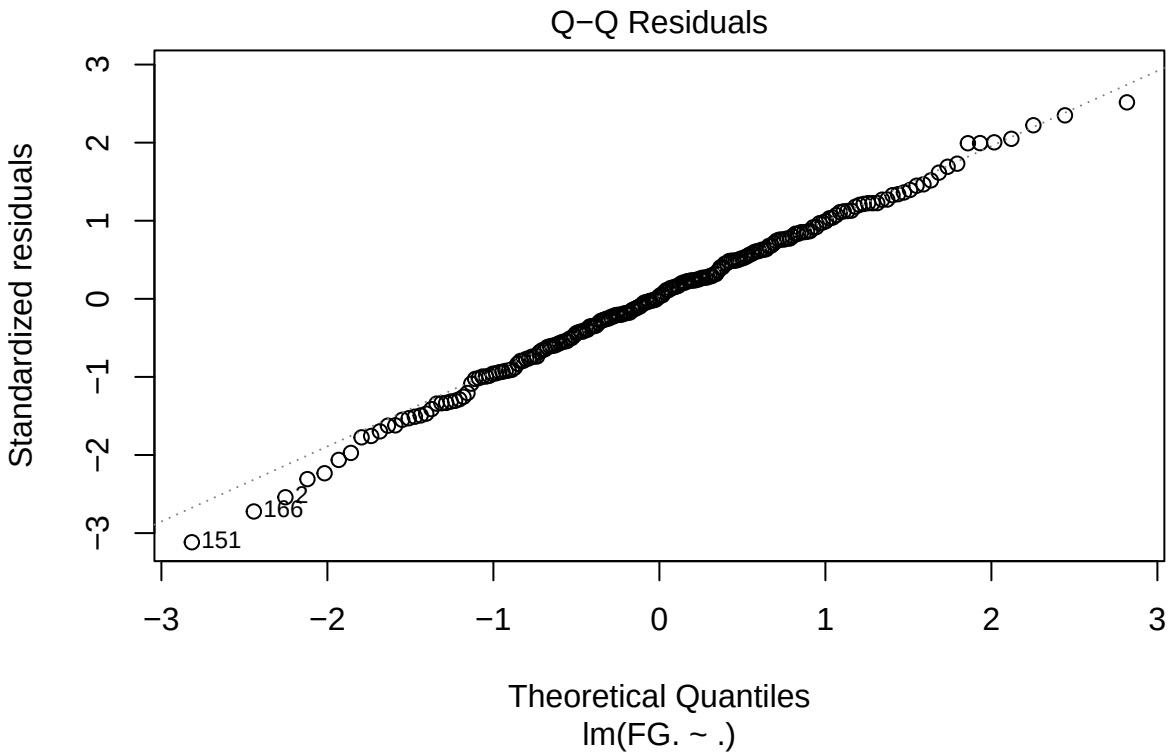
P-value 0.769 is not less than alpha 0.05 suggesting we cannot reject the null hypothesis and we conclude that residuals are normally distributed.

```
#Homoskedasticity
#Plot Fitted values on residuals
plot(lm_nba_norm, which=1)
```



Two ends look a little variable while most of the data is clumped around the middle.

```
#QQplot
plot(lm_nba_norm, which=2)
```



Data for the most part follows the linear aspect of the plot except the low end of the variables where it tails off a little.

```
#Remove outliers from previous two plots: 2, 151, and 166.
outliers <- stats_normal[-c(2,151,166),-6]
lm_outliers<- lm(formula = FG. ~ ., data = stats_normal[-c(2,151,166),-6])
summary(lm_outliers)
```

```
##
## Call:
## lm(formula = FG. ~ ., data = stats_normal[-c(2, 151, 166), -6])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52060 -0.43772  0.02647  0.43032  1.53409
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02770    0.04400   0.630   0.5297
## twoPR        0.52098    0.06959   7.486 2.24e-12 ***
## X3PA        -0.25384    0.05470  -4.640 6.30e-06 ***
## FTr          0.10374    0.05929   1.750   0.0817 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6268 on 199 degrees of freedom
## Multiple R-squared:  0.6043, Adjusted R-squared:  0.5983
```

```
## F-statistic: 101.3 on 3 and 199 DF,  p-value: < 2.2e-16
```

FTr is no longer significant so we need to remove it.

```
lm_outliers<- lm(formula = FG. ~ ., data = stats_normal[-c(2,151,166),-3])
summary(lm_outliers)
```

```
##
## Call:
## lm(formula = FG. ~ ., data = stats_normal[-c(2, 151, 166), -3])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46419 -0.44512 -0.02145  0.43474  1.55094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02886    0.04422   0.653    0.515
## twoPR        0.59804    0.05416  11.042  < 2e-16 ***
## X3PA        -0.23804    0.05423  -4.390 1.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.63 on 200 degrees of freedom
## Multiple R-squared:  0.5982, Adjusted R-squared:  0.5942
## F-statistic: 148.9 on 2 and 200 DF,  p-value: < 2.2e-16
```

```
#VIF of normalized data with no outliers and no non-significant predictors
vif(lm_outliers)
```

```
##    twoPR     X3PA
## 1.508059 1.508059
```

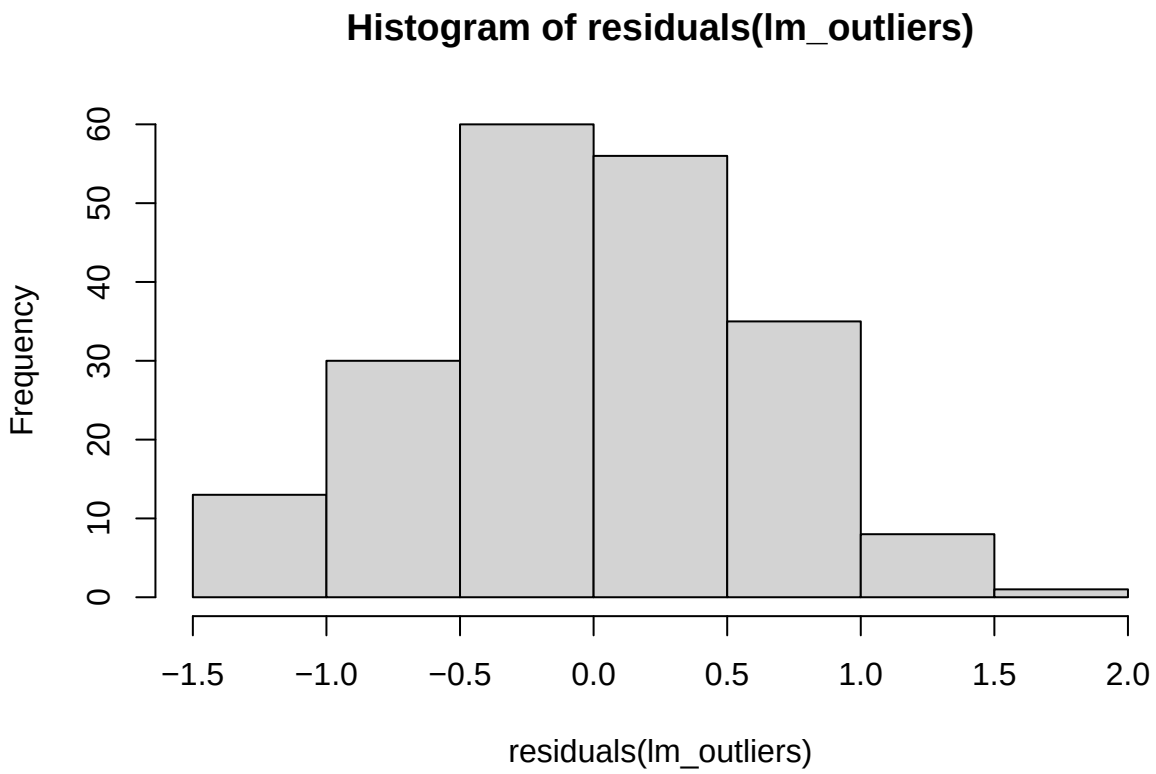
VIF is pretty low for both two percent rate and 3 point attempts.

```
#Shapiro test for normality of outliers
shapiro.test(residuals(lm_outliers))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm_outliers)
## W = 0.99452, p-value = 0.6674
```

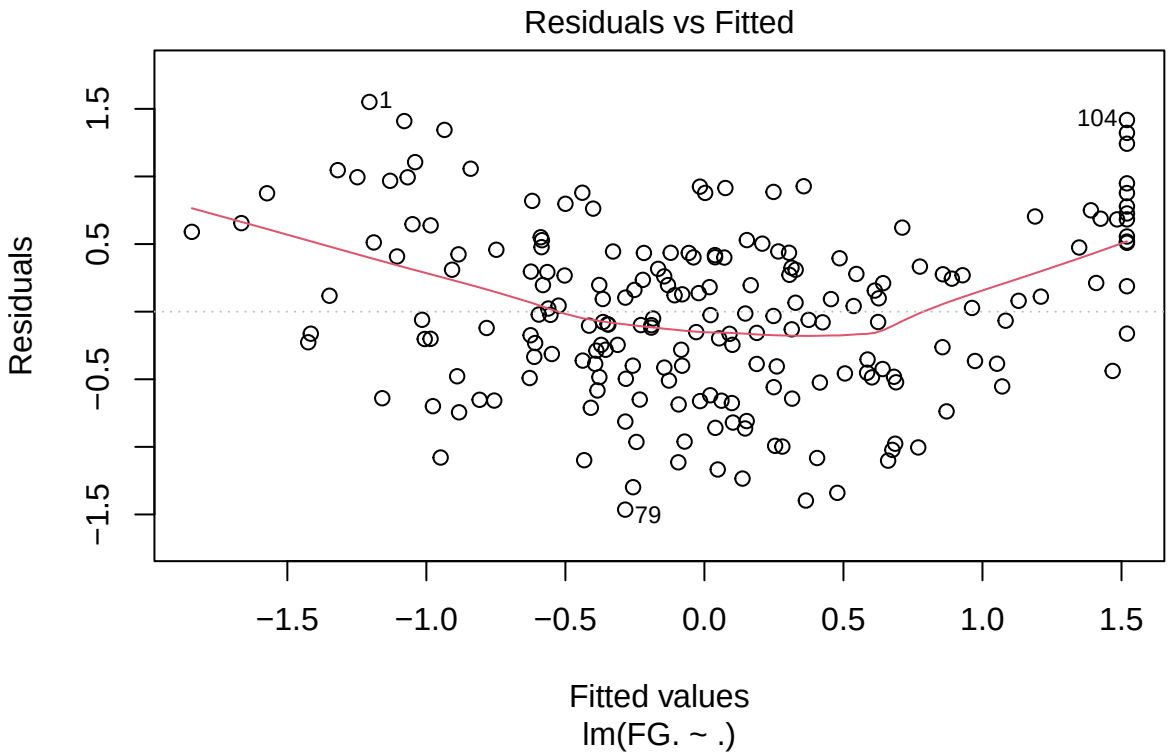
P-value 0.4772 is not less than alpha 0.05 suggesting we cannot reject the null hypothesis and we conclude that residuals are normally distributed.

```
#Histogram of residuals
hist(residuals(lm_outliers))
```



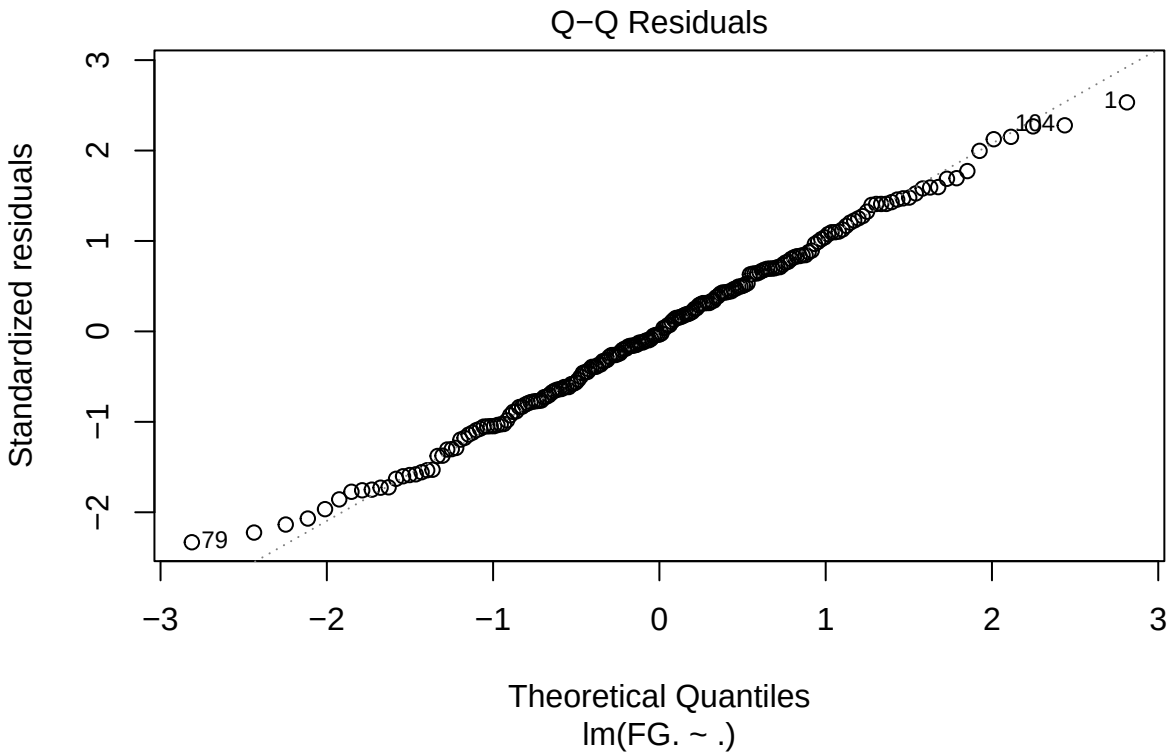
Looks normally distributed

```
#QQplot
plot(lm_outliers, which=1)
```



Looks a little better, but still not centered around 0 at the high and low ends.

```
#QQplot
plot(lm_outliers, which=2)
```



Follows a straight line for the most part.

```
# Set up data frame
final_df <- data.frame(actual = outliers$FG.,
           predicted = predict(lm_outliers))
```

```
#RMSE
sqrt(mean(residuals(lm_outliers)^2))
```

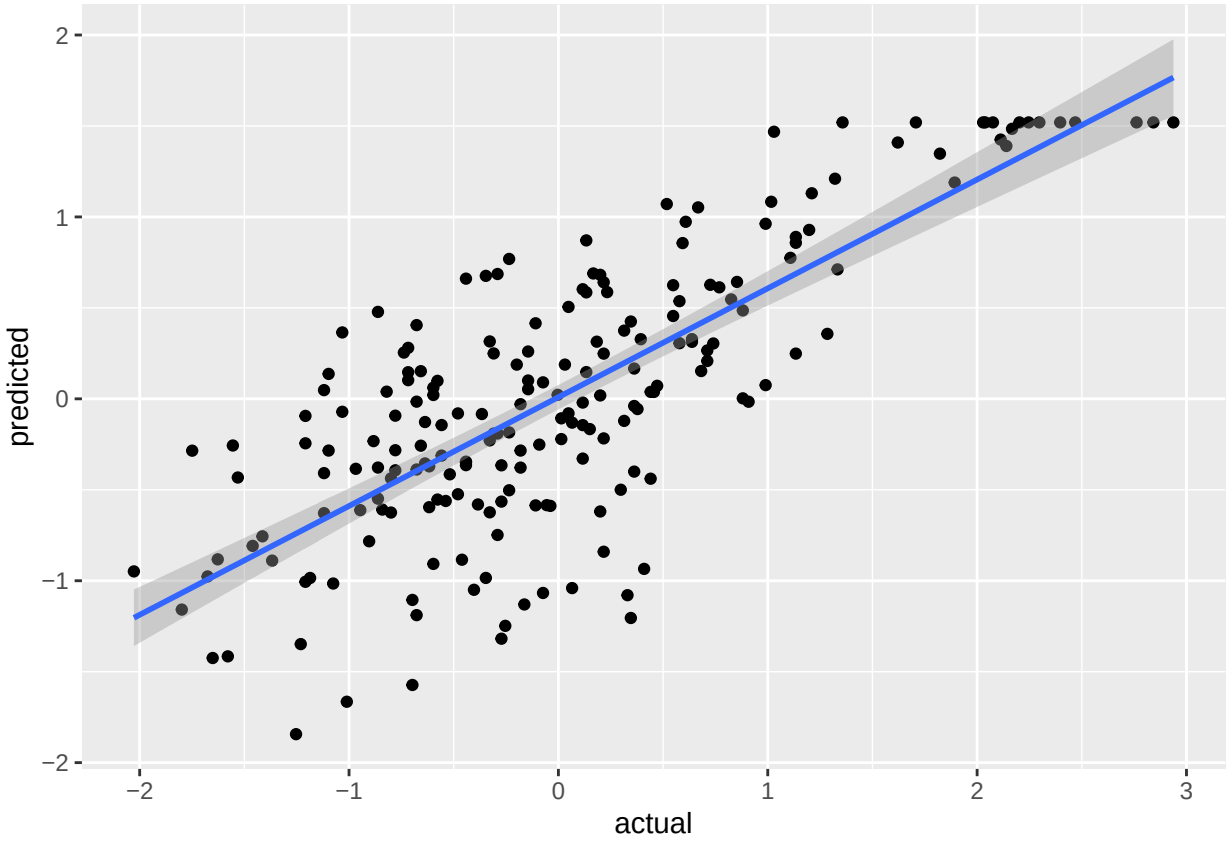
```
## [1] 0.6253545
```

```
#R^2
cor(outliers$FG.,
    predict(lm_outliers))^2
```

```
## [1] 0.5982225
```

```
# Plot fitted on actual
ggplot(data = final_df,
       aes(x = actual, y = predicted)) +
       geom_point() +
       geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Discussion

The initial question that was asked was if field goal percentage was related to any of the three variables: two point rate, three point attempts, and free throw rate. Initially looking at the data, two point rate seemed it would have the strongest relation with field goal percentage with a correlation of 0.761. When creating the linear model without transformations, each of the slopes are significantly different from 0 as they are 4.25e-13, 0.000139, and 0.03 respectively. Additionally, the R-squared of 0.6066 suggests that 60.66% of the variance of field goal percentage can be explained by two point rate, three point attempts, and free throw rate. Next, multicollinearity was investigated and the VIF of each of the factors was low suggesting multicollinearity was not a problem. Finally, the residuals were investigated and from the histogram it initially looked like the residuals could be normally distributed. However, the shapiro test not being significant in addition to the fitted values on residuals plot and qq plot indicated that the residuals were not normally distributed. As a result of this, analysis was done with the transformed data to see if results would be different.

In the linear model for the transformed data, each of the predictors are once again significant as the are less than alpha 0.05. Additionally, the r-squared value has gone down to 0.5607 meaning that 56.07% of the variance of the field goal percentage can be explained by two point rate, three point attempts, and free throw rate. For multicollinearity, the VIF of each of the predictors is low (all below 2.5) indicating multicollinearity is not an issue. For the normality of the residuals, the histogram appears to be normally distributed and this is further supported by the Shapiro-Wilk test where the p-value is 0.769 suggesting that the null hypothesis cannot be rejected and that the residuals are normally distributed. Additionally for the homoskedasticity, the values on the residual versus fitted plot display a major clump of point in the middle but variance on the two ends of the plot. Finally, for the qq-residual plot, it follows a straight line for the most part but tails off on the two ends. The outliers 151, 166, and 2 are highlighted so they are removed to create a new linear model.

In the new linear model without outliers, the free throw rate predictor became non-significant so that was removed. After that, the new linear model had an r-squared value of 0.5942 indicating 56.07% of the variance of the field goal percentage can be explained by two point rate and three point attempts. The VIF of the new model is low with both predictors being around 1.5 meaning multicollinearity is again not a problem. The shapiro test (p-value above 0.05 and at 0.6674) and histogram both indicate that the residuals are normally distributed. Additionally, the homoskedastcity looks solid as the residuals vs fitted and the qq plots both follow as they should indicating normality of the residuals. This was the final model. From this, the B1 for two point rate of 0.598 and p value being less than 0.01 indicated that with each standard deviation increase in two point rate, there is a 0.598 standard deviation increase in field goal percentage when controlling for three point attempts. Additionally, because it is statistically significant, B1 for two point rate is likely not equal to zero. Next, the B2 of three point attempts being -0.238 and the p value being less than 0.01 indicated that with each standard deviation increase in three point attempts, there was a 0.238 standard deviation decrease in field goal percentage. The p value also suggests that three point attempts are likely not equal to zero. The RMSE of the final model is 0.625 indicating that on average, our predicted values are about 0.625 standard deviations away from the actual values. The R^2 value of 0.598 suggests that 59.8% of field goal percentage can accounted for by two point rate and three point attempts.

Overall, we can conclude that the final model does not do a great job predicting field goal percentage based off of the variables two point rate, three point attempts, and free throw rate. In most situations, the R^2 is not very high. Additionally, without transforming the data, the residuals are not normally distributed Even when the data is normalized, and the residuals are normal the RMSE is quite high and the R^2 is also quite low.