# Debunking NBA Statistics

06.07.2020

Justin Lee
University of California Santa Barbara
PSTAT 126 - Regressional Analysis
Professor Sudeep Bapat

## Overview

As an avid fan of the Lakers and the organization itself, I could not pass the opportunity to attempt the final project with the NBA statistics to perform regressional analysis. Sports betting has been around for many decades whether it is done illegally on a website or in your house with your friends. How hard is it to guess the team that will take home the championship? As of now, the NBA website provides the public with statistics that are accumulated from the hardworking data scientists in the NBA.  An important aspect of this research is that the NBA culture that we are in now is an era of three-point shooting. Steph Curry of the Golden State Warriors has changed the culture of the game by shifting the scoring method from midrange to three-pointers.

This leads us to my purpose, how can we use regressional analysis to provide a statistic-based estimate of which team will take the championship home?

The thesis will analyze correlations between the team's  win/loss statistics along with the average three-pointers. The statistics were provided by the NBA website and I will be using the 2018-2019 regular season statistics per team to see whether if I could correctly see if this method will be accurate.

## Question of Interest

Are the number of wins determined by average points, FGM, FG%, 3PM, and 3P%?

## Regression Method

As stated above, the statistics for the NBA is given from *www.stats.nba.com*.  To assure the accuracy of the model, I chose the most recent year where the NBA season finished which is 2018-2019. Luckily for me, the site provided extensive amounts of data and pretty much every single statistic of the games played per team.  From this information, I will be building a model to meet the four LINE conditions. This will ensure the most important predictors in the model.

## Variables

- Y = W\L (Win/Loss)
- X1 = PTS(Points)

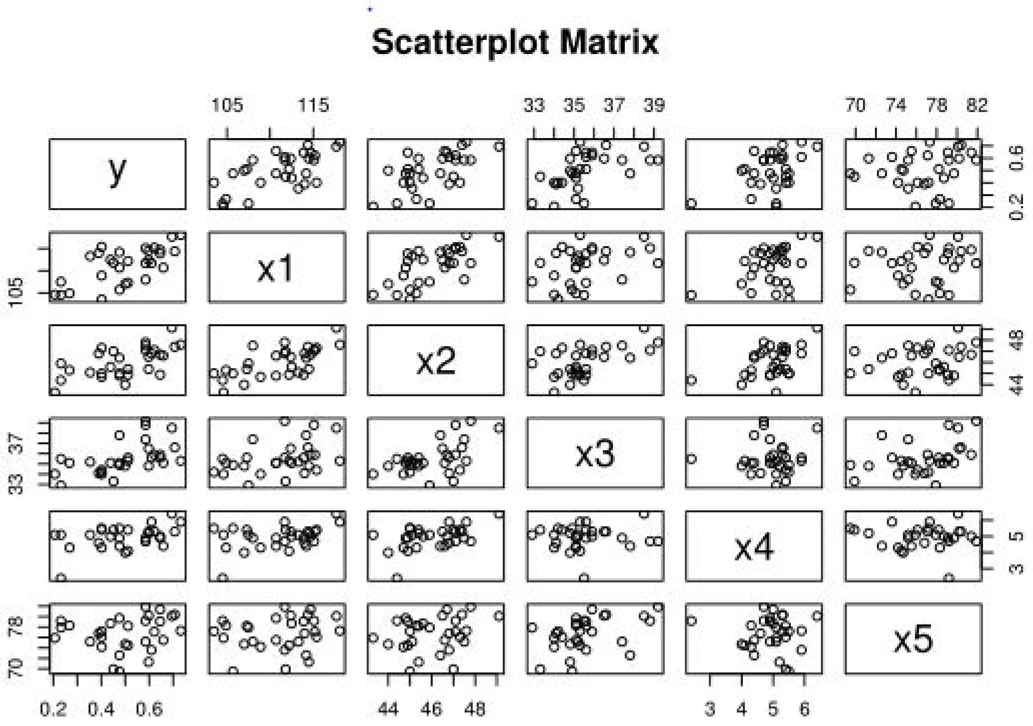
- X2 = FG%(Field goals made/field goals attempted)
- X3 = 3P%(3 points made/ 3 points attempted)
- X4 = TOV(turnovers)
- X5 = FT%(Freethrows made/ freethrows attempted)

*It is important to note that these are the averages per season*

Y is the variable we are trying to predict with the predictor variables x1,x2,...,x5.

## Regression Analysis, Results, and Interpretation

First I used the pairs() function in R to see the plot.



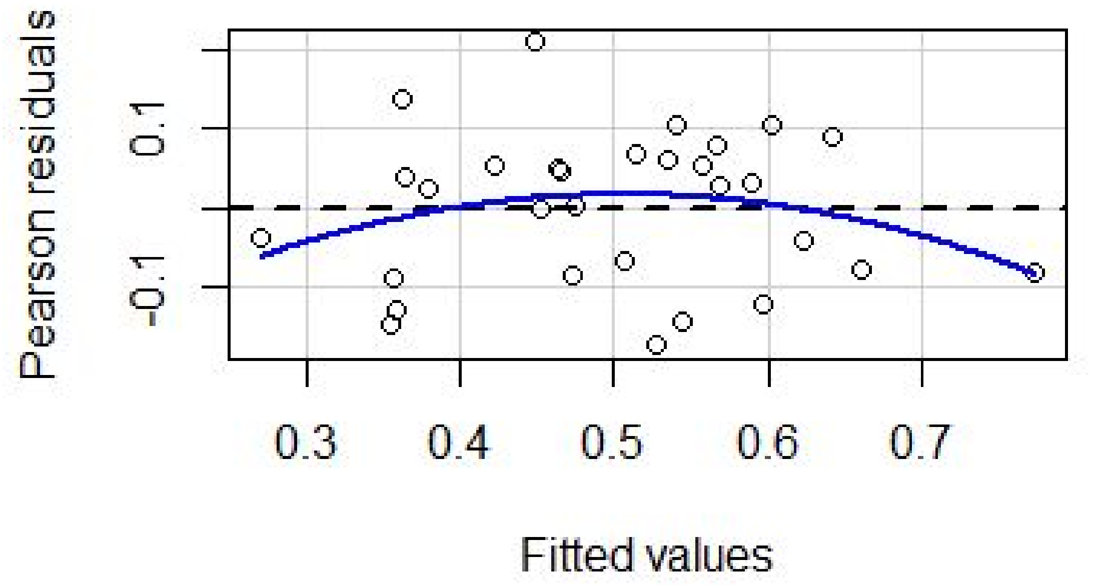
Now it seems appropriate to use the stepwise function [step()] to decide which predictor variables will have the lowest AIC score which ensures accuracy in our analysis.

```
## Start: AIC=-114.18
## y ~ 1
##
## Df Sum of Sq RSS AIC
## + x1 1 0.272538 0.35163 -129.39
## + x2 1 0.232651 0.39151 -126.17
## + x3 1 0.183083 0.44108 -122.59
## + x4 1 0.118108 0.50606 -118.47
## <none> 0.62416 -114.18
## + x5 1 0.016084 0.60808 -112.96
##
## Step: AIC=-129.39
## y ~ x1
##
## Df Sum of Sq RSS AIC
## + x3 1 0.060554 0.29107 -133.06
## + x2 1 0.031622 0.32000 -130.22
## <none> 0.35163 -129.39
## + x4 1 0.021438 0.33019 -129.28
## + x5 1 0.002017 0.34961 -127.56
## - x1 1 0.272538 0.62416 -114.18
##
## Step: AIC=-133.06
## y ~ x1 + x3
##
## Df Sum of Sq RSS AIC
## + x4 1 0.032167 0.25891 -134.57
## <none> 0.29107 -133.06
## + x2 1 0.006768 0.28430 -131.77
## + x5 1 0.003158 0.28792 -131.39
## - x3 1 0.060554 0.35163 -129.39
## - x1 1 0.150009 0.44108 -122.59
##
## Step: AIC=-134.57
## y ~ x1 + x3 + x4
##
## Df Sum of Sq RSS AIC
## <none> 0.25891 -134.57
## - x4 1 0.032167 0.29107 -133.06
## + x5 1 0.001685 0.25722 -132.77
## + x2 1 0.000103 0.25880 -132.59
## - x3 1 0.071282 0.33019 -129.28
## - x1 1 0.077052 0.33596 -128.76
##
## Call:
## lm(formula = y ~ x1 + x3 + x4)
##
## Coefficients:
3
## (Intercept) x1 x3 x4
## -2.67318 0.01500 0.03534 0.05022
```

From this table, we can see that the best fitting model will be a regression model using x1,x3, and x4. X1 is the average of points scored per season, x3 is the three-point percent average, and x4 is the average number of turnovers. This step gives us an AIC value of -134.57 which is the lowest. It is also important to note that when it comes to AIC, we do not need to find the value closest to zero, but the number that is the lowest whether it is negative or positive.

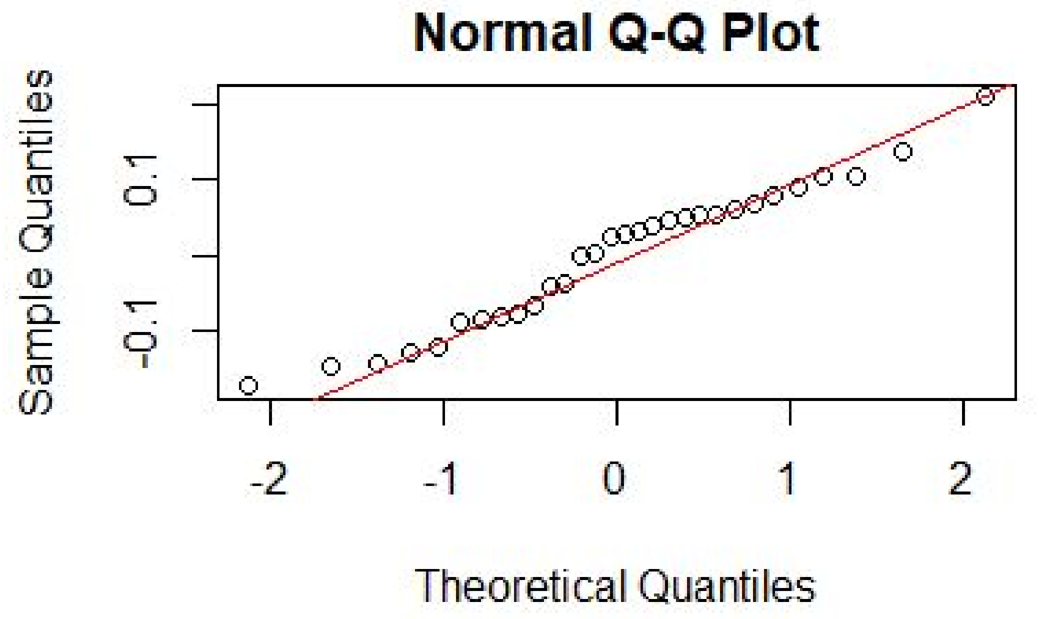
```
## Call:
## lm(formula = y ~ x1 + x3 + x4)
##
## Coefficients:
## (Intercept) x1 x3 x4
## -2.67318 0.01500 0.03534 0.05022
```

Now to check the L.I.N.E assumption we must first check the first step, **Linearity.** I will be drawing a scatter plot of residuals and y values using the residualPlot() function.

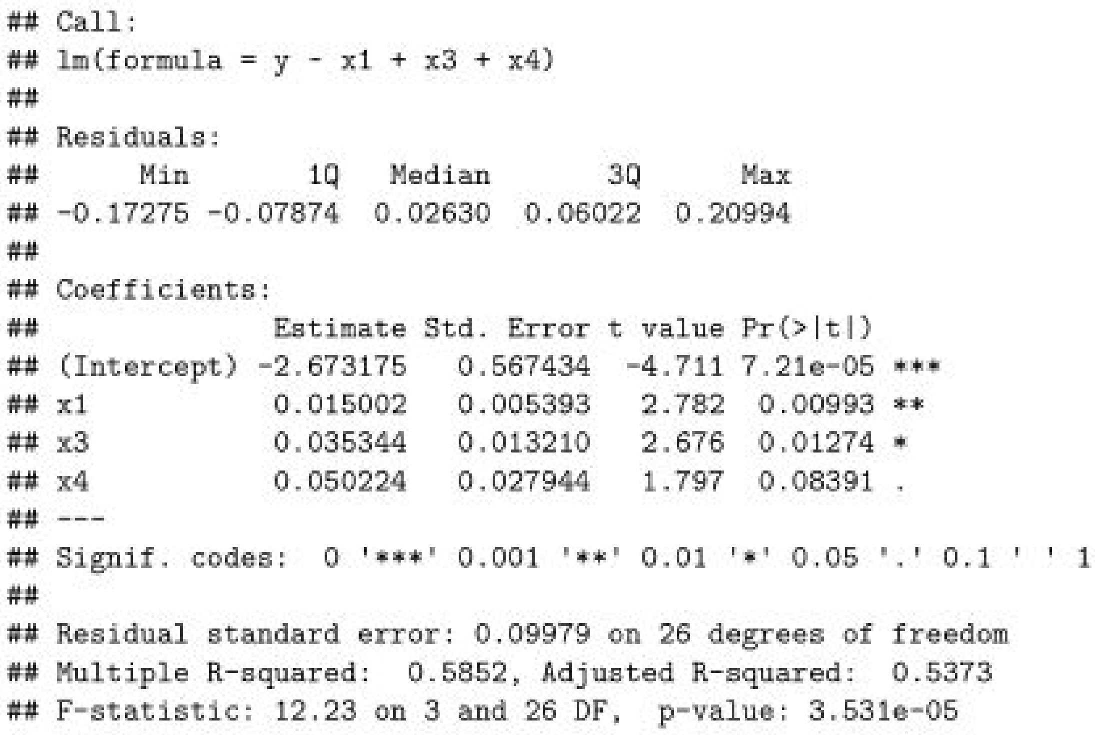


As we can see from the graph, we could see that the blue line is fairly flat and the points scatter pretty evenly which means we could assume linearity in this example.

Now to test for normality, I will be using a normal Q-Q plot.



From this plot, we can see that the plot follows a linear pattern along the line which is evidence to say that our normality condition is met.



```
## Call:
## lm(formula = y ~ x1 + x3 + x4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17275 -0.07874  0.02630  0.06022  0.20994
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.673175   0.567434  -4.711 7.21e-05 ***
## x1           0.015002   0.005393   2.782  0.00993 **
## x3           0.035344   0.013210   2.676  0.01274 *
## x4           0.050224   0.027944   1.797  0.08391 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09979 on 26 degrees of freedom
## Multiple R-squared:  0.5852, Adjusted R-squared:  0.5373
## F-statistic: 12.23 on 3 and 26 DF,  p-value: 3.531e-05
```

To summarize the table above, we can see that 58.52% of all variations of the w/l ratio is explained by the average of points scored per season, three-point percent average, and the average number of turnovers. The p-value is extremely small which does mean that our model does fit our hypothesis and it does confirm the original hypothesis that these factors do have an impact on whether a team has a high win/loss ratio.  Also, the F-statistic is 12.23 on 3 and 26

## Conclusion

From the analysis done, it is fair to conclude that the factors that contribute most to winning games in the NBA are the three points made, turnovers, and field goals made. The turnover numbers seem to play the biggest role in the number of wins.

It is also very important to realize that although these statistics show that these factors are important there are many other aspects that go into the teams winning such as player's health, how many Allstars there are, the potential of young players, etc… From my calculations, it was suspected that the Bucks will win the championship, but as seen in the last playoff, Kawhi Leonard of the Raptors went off and led them to win the championship.

I also do firmly believe that the three points made per team have a huge impact on whether the team has a high win/loss ratio or wins the championship because since this is a new style of playing, most teams do not really know how to defend from that.

## Appendix

```
library(readr)
library(tidyverse)
library(alr4)
library(lindia)
nba <- read_csv("~/school/3. Spring 2020/PSTAT 126 (real)/final project/nba/nba 2018-2019.csv")

y = nba$`W/L`
x1 = nba$PTS
x2 = nba$`FG%`
x3 = nba$`3P%`
x4 = nba$TOV
x5 = nba$`FT%`
pairs(~y+x1+x2+x3+x4+x5,data = nba, main = "Scatterplot Matrix")

mod0 <- lm(y~1)
```

```
mod.all<- lm(y~x1+x2+x3+x4+x5)
step(mod0, scope = list(lower = mod0, upper= mod.all))

fit <- lm(y~x1+x3+x4)
summary(fit)
res = resid(fit)
y_hat = fitted(fit)
plot(y_hat, res, main = "Residual vs. Fit", ylab = "Residuals", xlab = "Fit")
abline(h = 0, lty = 2)

residualPlot(fit)

qqnorm(res)
qqline(res, col = 2)

anova(fit)
summary(fit)
```

## Work Cited

"Teams Traditional." *NBA Stats*, stats.nba.com/teams/traditional/?sort=W_PCT&dir=-1&Season=2018-19&SeasonType=Regular%2BSeason.