Analyzing the factors that affect wins in sports game

Dongkyu Kim

2020 12 10

Code and data supporting this analysis is available at (https://github.com/DongkyuKim-max/stat304)

Abstract

The purpose of this study is to see if there is any relationship between a team's players' average age, payroll, moving distance, average points scored and the numbers of winnings in the basketball game. The two regression analysis is done to identify the relationship. The first method used was simple linear regression. There was a strongly positive relationship between players' payroll and average points scored with the number of wins. The more and higher payroll and average points teams got during the season, the higher number of wins. In addition, the multiple linear regression model was used to identify what are the factors that are not crucial for the winning. As a result, the average age of team and moving distance were considered as not important factor. The results have significant implications for determining the number of wins.

Keywords: Simple linear regression, sports game, NBA, Multiple linear regression

Introduction

There are different types of variables that affect the sports game. As a student who likes to play basketball, I always had concerns what are the most important factors that lead to higher winning ratio. The possible factors are players payroll, ages, heights, number of passing etc. There are many more that one need to consider when they are building the team. The use of statistical analysis of basketball games using multiple linear regression model, simple linear regression and multilevel modeling. In addition, through this model, one could determine which conference gives more probability to win the first place. For this model, payroll is selected as one of the important variable, because many people believe that scouting players with high payroll will make it easier to win. Furthermore, average age of players is another important factor that can affect players' stamina and performance. Lastly, the average points per game is also another good indicator of the standing. The team with high payroll, younger average age and the high average points per game lead to the winning of the league.

Data

The data is selected from NBA website. The data on the website is categorized by the season. By yearly. Among the data, the data from 2018-2019 was selected as it is considered to be a most recent data. As there were tremendous data about the team and players, only several variables were selected. The chosen variables are Age, Payroll of the players, moving distance by the team, number of wins and average points by each

team. The selected data were downloaded and customized and ranked by the standings and the conference during the regular season. The values are all numerical. From the official NBA website, all the data were collected by hand and saved in the excel file.

Model

The model focused on major factors that affect the number of wins in basketball game. Simple linear regression and multiple linear regression methods were used to identify the factors. To begin with, from the data selected. The chosen variables were average payroll of the players, average age of the players and average points each team scored during the regular season. The simple linear regression gives a positively correlated relationship with independent variables. In addition, multiple linear regression method was used to identify the how chosen variables predict the outcome of the number of wins.

Results

Figure 1 shows the simple linear regression model of number of wins and average players payroll. It has positively correlated relationship. The P-value of 0.05 indicates that there is a significant difference between the variables. As shown in Figure 1, the higher average payroll team pays to the players, the higher number of wins they get.
Figure 2 shows the simple linear regression model of number of wins and average points the team got. It has positively correlated relationship. The P-value of 0.000358 indicates that there is a significant difference between the variables. As shown in Figure 2, the higher points they made in each game leads to higher number of wins during the season.
Figure 3 shows the simple linear regression model of number of wins and average ages of players in the team. It has positively correlated relationship. The P-value of 0.012 indicates that there is a significant difference between the variables.It was expected to have a negatively correlated relationship between the independent and dependent variables. However, as shown in figure 3, it is positively correlated. It can rejects the hypothesis, which states team with younger average age leads to higher number of wins. Multiple linear regression method was done to show the effect of different variables on number of wins. As shown in figure 4, the p-value of 0.0001961 shows that variables provide enough evidence to support the original hypothesis. In addition, the Adjusted R-squared value of 0.5047 also supports the hypothesis. However, On the one hand, the p-value for both payroll and average point suggest that these two factors have really huge impact on the dependent variables. On the other hand, the p-value for average age and travel distance indicate that these two factors are not reliable. The graph of Residuals vs Fitted is used to identify the non-linearity and outliers. One can identify that the red line is closed to the dashed line, which means the linearity of the graph seems to hold. In addition, the normal Q-Q plot demonstrates the two sets of quantiles against each other. The graph demonstrates roughly straight line. This indicates both sets are came from normal distribution.

Discussion

There are several weaknesses in this analysis. One of the weaknesses is the lack of numbers of variables in the data set. The only dataset has is information about NBA season during 2018-2019 season. In order to have a better determination of the winning, it is important to have a large amount of information in order to generalize the conclusion. Furthermore, lack of categorical values is another weakness and should be improved for deeper analysis. Even though, the analysis focuses on factors that affect the number of winnings, it has to focus more on numerical values. However, I believe categorical values are also could be the important factors that contribute to number of wins. Lastly, the comparison between the used dependent variables were not discussed deeply in this analysis. For example, the results have shown that variables such

as players' average payroll and and average points scored showed a direct positive relationship with the number of wins. However, in this analysis there is not comparison between these two variables. For example, identifying and determining the most important factor could be one of the possible improvement. For next step, even though the analysis has been done, there are several improvements could be done for the future. In this paper, simple linear regression model and multiple linear regression are used with the data collected. However, there are several different statistical model that could be also used such as logistic function, post-stratification and Bayesian. The use of these functions could give better results for the aim of this project. As stated in the weakness, all the data selected were quantitative, it limits the use of logit function. The analysis will be significantly improved if it contains categorical values and the analysis.

Appendix

```
library(tidyverse)
```

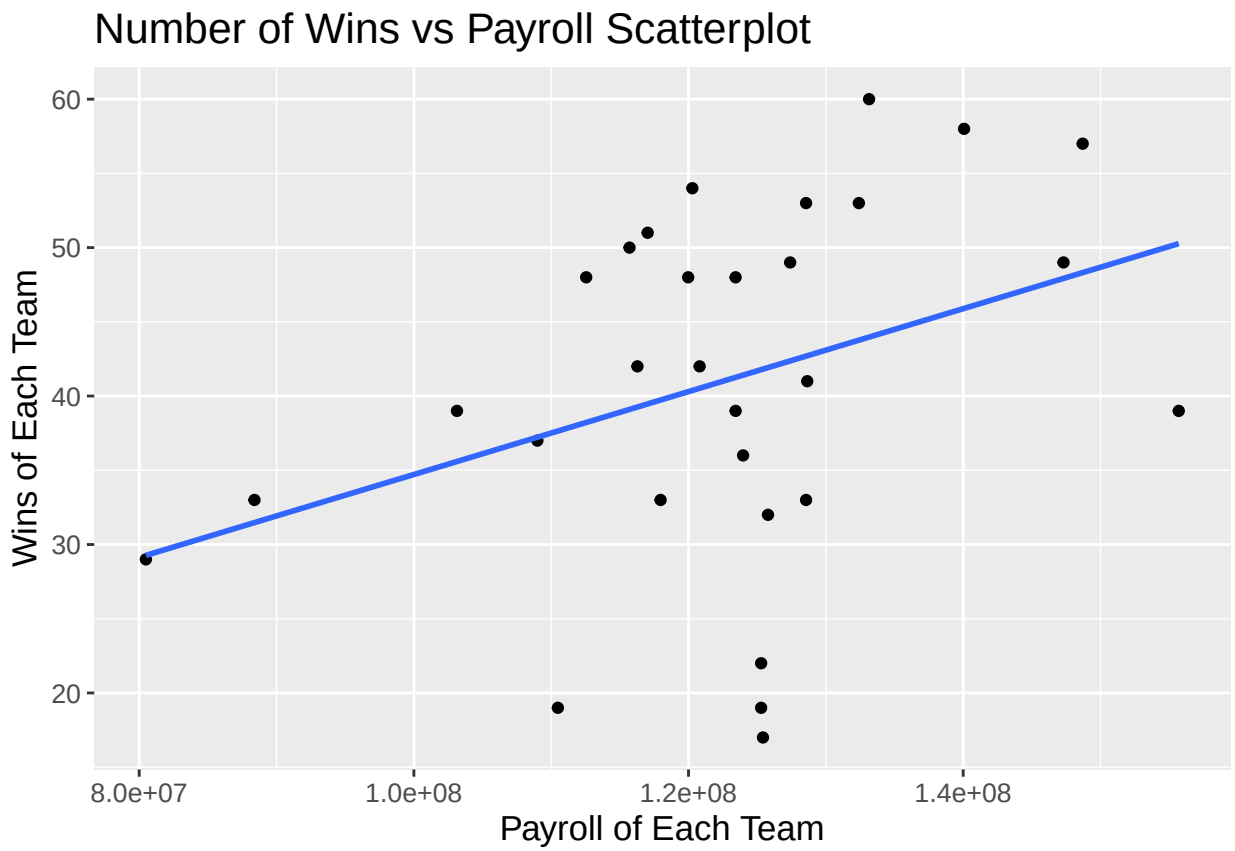
```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(readxl)
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.0.3

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
STAT_304_data<-read_xlsx("C:/Users/PC/Desktop/stat project/STAT_304_data.xlsx")
```

```
#Figure 1.Simple linear regression model of wins and payroll.
ggplot(data = STAT_304_data, ) +
  aes(x = Payroll, y = Win) +
  geom_point() +
  theme(text = element_text(size=20)) +
  ggtitle("Number of Wins vs Payroll Scatterplot") +
  theme(text = element_text(size = 13)) +
  labs(y = "Wins of Each Team", x= "Payroll of Each Team") +
  geom_smooth(method='lm', formula = y~x, se = FALSE)
```

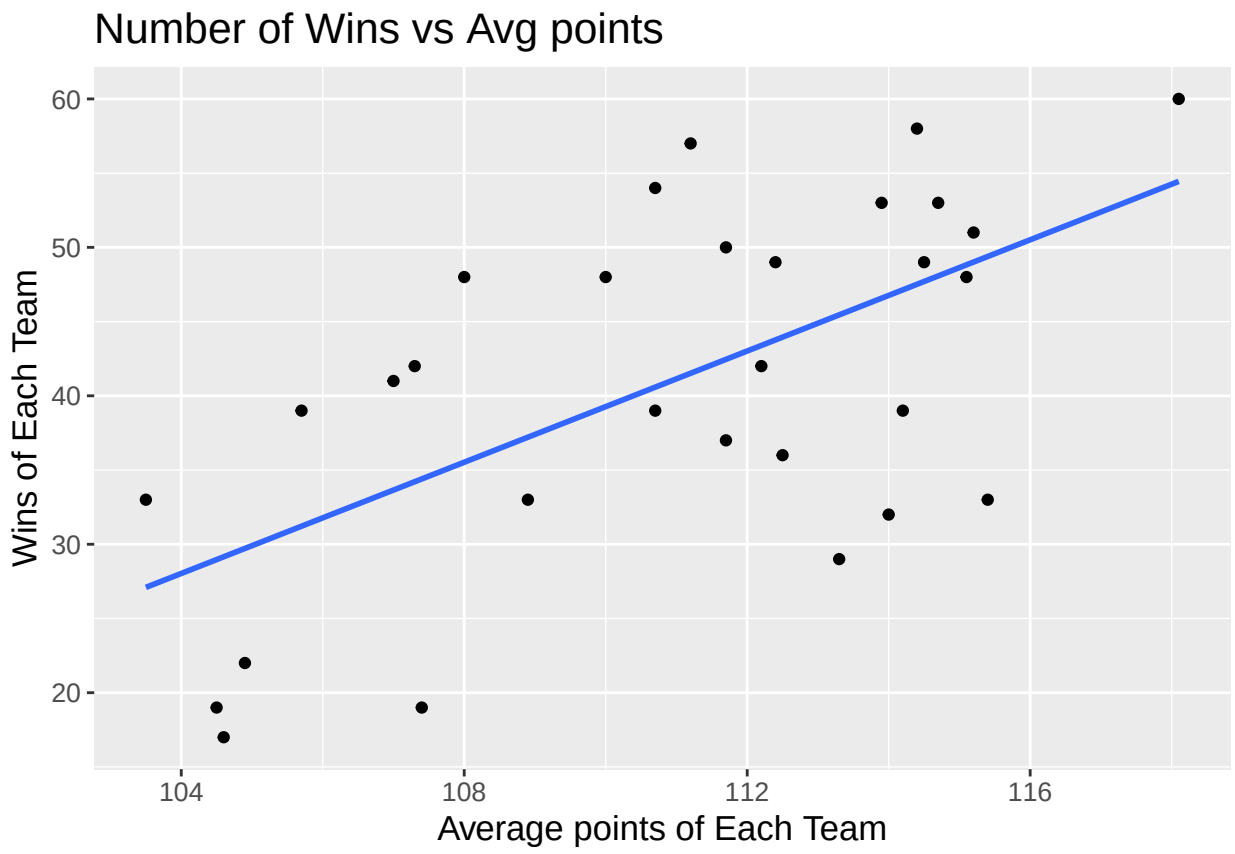



```
summary(lm(STAT_304_data$Win~STAT_304_data$Payroll, data = STAT_304_data))
```

```
##
## Call:
## lm(formula = STAT_304_data$Win ~ STAT_304_data$Payroll, data = STAT_304_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.814  -6.402   1.500   9.098  16.028
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.792e+00  1.688e+01   0.402   0.6905
## STAT_304_data$Payroll 2.792e-07  1.367e-07   2.042   0.0506 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.42 on 28 degrees of freedom
## Multiple R-squared:  0.1296, Adjusted R-squared:  0.09856
## F-statistic: 4.171 on 1 and 28 DF,  p-value: 0.05064
```

```
#Figure 2. Simple linear regression model of wins and Average points

ggplot(data = STAT_304_data, ) +
  aes(x = Average_point, y = Win) +
```

```
  geom_point() +
  theme(text = element_text(size=20)) +
  ggtitle("Number of Wins vs Avg points") +
  theme(text = element_text(size = 13)) +
  labs(y = "Wins of Each Team", x= "Average points of Each Team") +
  geom_smooth(method='lm', formula = y~x, se = FALSE)
```

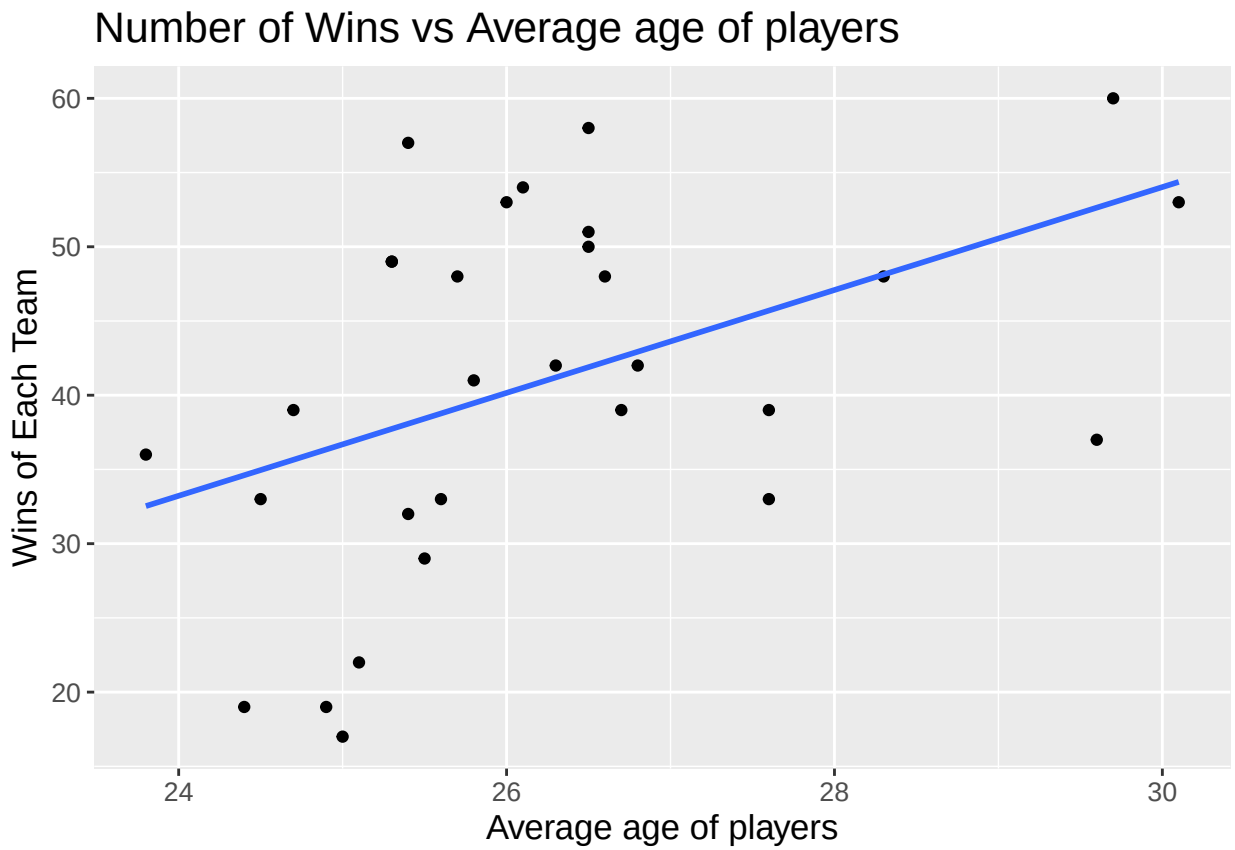



```
summary(lm(STAT_304_data$Win~STAT_304_data$Average_point, data = STAT_304_data))
```

```
##
## Call:
## lm(formula = STAT_304_data$Win ~ STAT_304_data$Average_point,
##     data = STAT_304_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.452  -7.894   1.645   7.496  15.482
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -166.7777    51.2057  -3.257 0.002946 **
## STAT_304_data$Average_point    1.8732     0.4614   4.060 0.000358 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 9.712 on 28 degrees of freedom
## Multiple R-squared:  0.3706, Adjusted R-squared:  0.3481
## F-statistic: 16.48 on 1 and 28 DF,  p-value: 0.0003576
```

```
#Figure 3. Simple linear regression model of wins and average age of players

ggplot(data = STAT_304_data, ) +
  aes(x =Age, y = Win) +
  geom_point() +
  theme(text = element_text(size=20)) +
  ggtitle("Number of Wins vs Average age of players") +
  theme(text = element_text(size = 13)) +
  labs(y = "Wins of Each Team", x= "Average age of players") +
  geom_smooth(method='lm', formula = y~x, se = FALSE)
```




```
summary(lm(STAT_304_data$Win~STAT_304_data$Age, data = STAT_304_data))
```

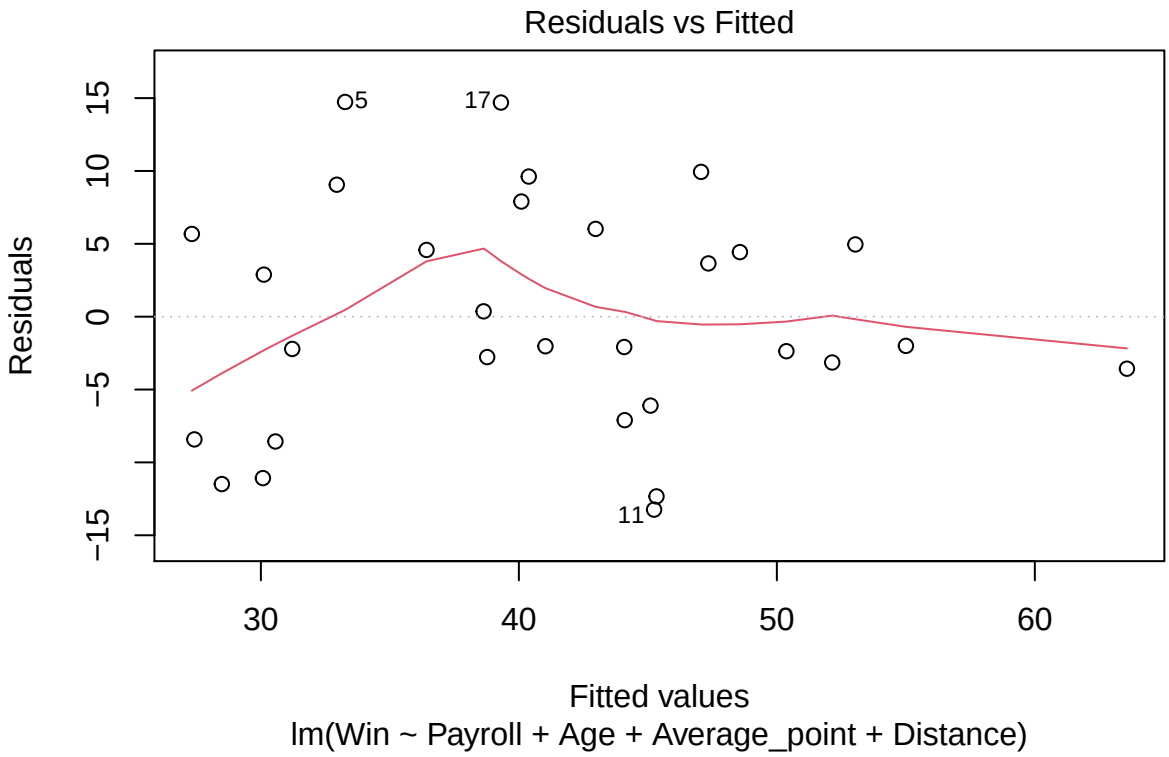
```
##
## Call:
## lm(formula = STAT_304_data$Win ~ STAT_304_data$Age, data = STAT_304_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6907  -6.5459   0.3376   8.6900  18.9230
##
```

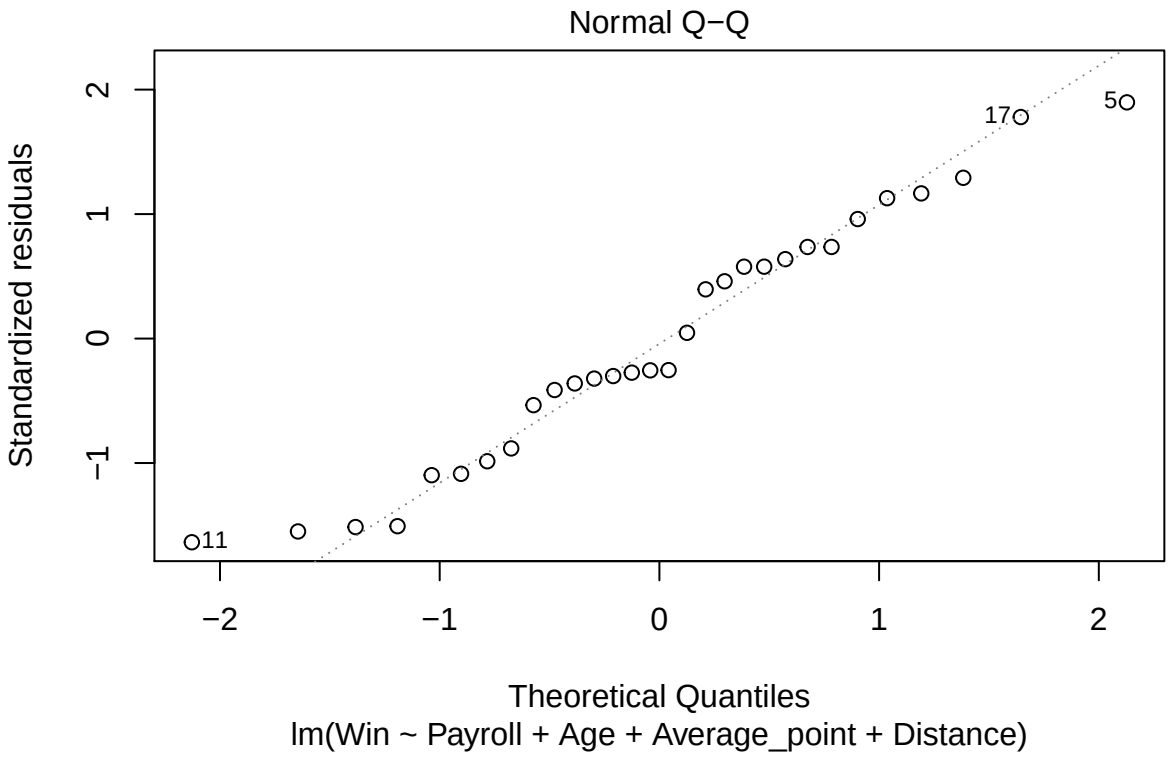
```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -49.958     34.080  -1.466   0.1538
## STAT_304_data$Age    3.466      1.296   2.674   0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.93 on 28 degrees of freedom
## Multiple R-squared:  0.2034, Adjusted R-squared:  0.1749
## F-statistic: 7.148 on 1 and 28 DF,  p-value: 0.01238
```

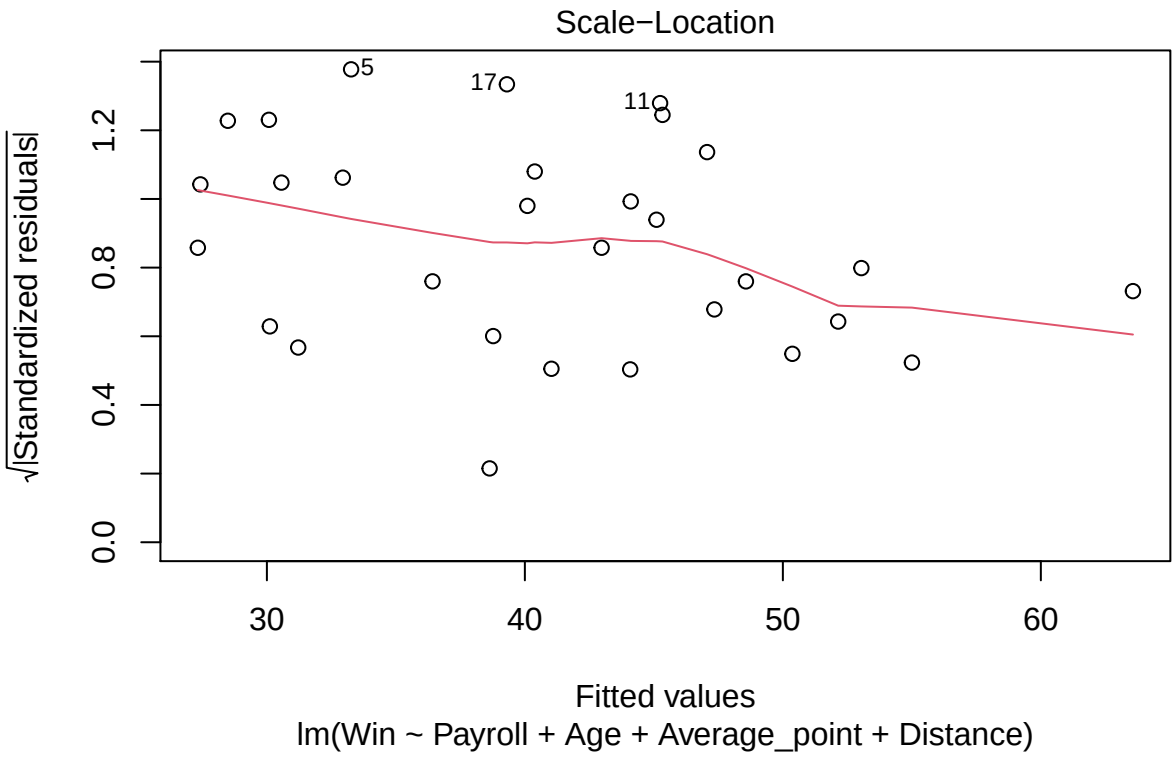
```
#Figure 4. Multiple linear regression model of wins and all factors.
Multiple_linear_Regression <- lm(formula = Win ~ Payroll +Age +Average_point+Distance, data = STAT_304_

summary(Multiple_linear_Regression)
```

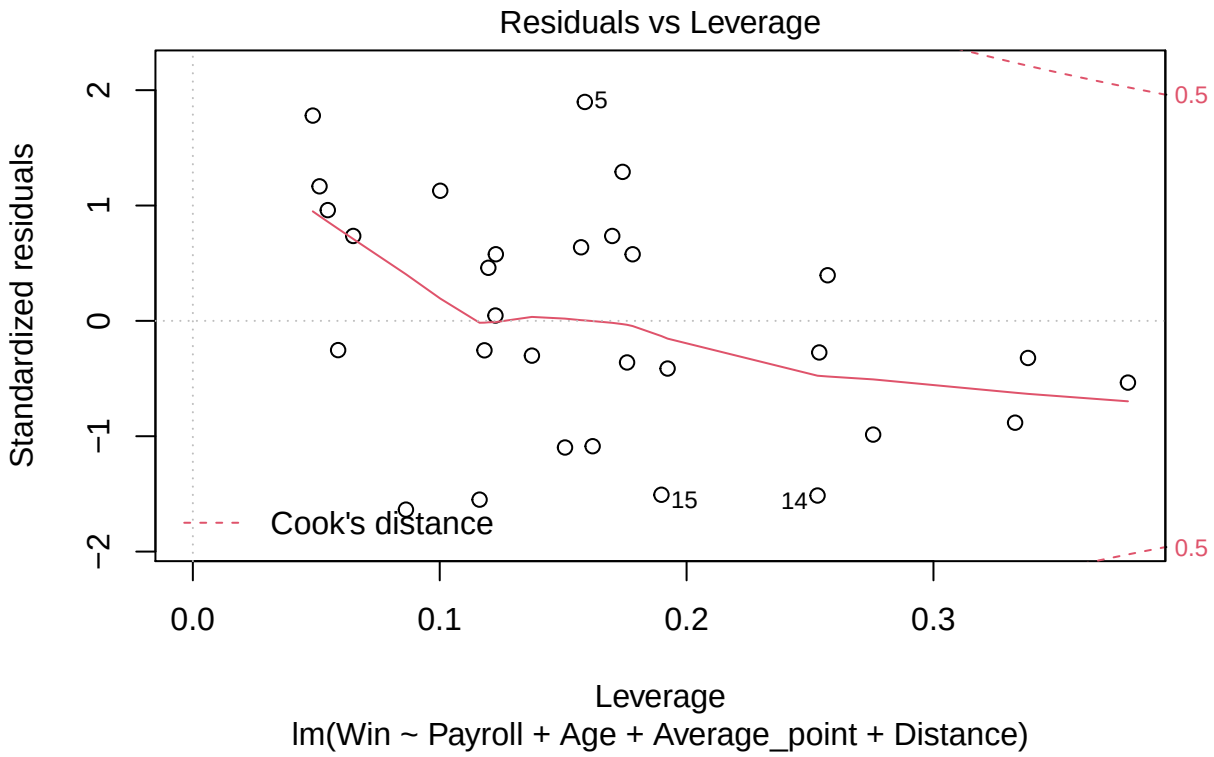
```
##
## Call:
## lm(formula = Win ~ Payroll + Age + Average_point + Distance,
##     data = STAT_304_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.242  -5.469  -2.017   5.496  14.735
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.211e+02  4.853e+01  -4.556 0.000118 ***
## Payroll        3.009e-07  1.015e-07   2.964 0.006585 **
## Age            2.019e+00  1.107e+00   1.824 0.080093 .
## Average_point  1.603e+00  4.390e-01   3.651 0.001208 **
## Distance      -1.192e-04  3.082e-04  -0.387 0.702220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.466 on 25 degrees of freedom
## Multiple R-squared:  0.573,  Adjusted R-squared:  0.5047
## F-statistic: 8.386 on 4 and 25 DF,  p-value: 0.0001961
```

```
plot(Multiple_linear_Regression)
```


Residuals vs Fitted
Residuals
15
10
5
0
−5
−15
5
17
11
30
40
50
60
Fitted values
lm(Win ~ Payroll + Age + Average_point + Distance)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Win ~ Payroll + Age + Average_point + Distance)
11
17
5
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
5
17
11
30
40
50
60
Fitted values
lm(Win ~ Payroll + Age + Average_point + Distance)



References

NBA. "NBA.com/Stats." NBA Stats, www.nba.com/stats/.

"These Are the Salaries of All NBA Teams." HoopsHype, HoopsHype, hoopshype.com/salaries/.

Willman, Daren. "NBAsavant.com." NBAsavant.com: Your Source For Advanced NBA Analytics, nbasavant.com/apps/map.php.