What is the Recent Trend of Soccer in the Top Soccer Leagues

Eung Kyu Kim

December 20, 2020

Github Repository:

https://github.com/williamkim1102/Recent-Trend-of-Soccer

Abstract

In this report, an analysis will be performed to find the recent trend of soccer in the German, English and Spanish Leagues from 2019-2020. Variables related to soccer will be used in linear regression models to find the relationship between the variables. From these models, a recommendation of how soccer should be played nowadays will be given. The data are from the official websites of the three leagues.

Keywords


```
+ Soccer Trend
+ Data Analysis
+ Linear Regression
+ Tactics
+ Multiple Linear Regression
+ How to earn points in soccer
```


Introduction

Each country in the world has different opinions and political views, so the countries have been fighting against each other. Then what is the best thing that unites the world? One of the best ways to unite the world is through sports, especially, football. Football, as known as soccer in North America, is the most popular sport in the world and every country plays soccer. However, soccer is one of the sports that is extremely hard to predict the outcomes, which means there are lots of upsets and uncertainties. Recently, famous soccer teams' revenues have grown, so they have been spending more money to buy players (Connelly, 2020). However, soccer is not always about money and there are many other factors that can affect the outcomes.

The trends of soccer has been constantly changing. In the 1970s, a tactic called 'Total Football' was played by the Dutch soccer team. This tactic was all about adopting the roles of any other player in the team. (Siregar, 2018) This tactic required constantly running around the field, which means the players had to cover a lot of distance and it required stamina. After decades, around the early 2010s, a tactic called 'Tiki Taka' was the trend of soccer. Tiki Taka was used by the Spanish League team, FC Barcelona and the Spanish National Team and it was all about short passing and possession (Bairner, 2020). This tactic was one of the most influential tactic in soccer history and this made a lot of soccer teams to play with having high possession.

The most recent influential tactic of soccer was played by an English team called Leicester City FC in 2016. Their tactic also completely changed the way of thinking about soccer. Due to the influence of Tiki Taka, in the middle of 2010s, almost all the teams tried to have high possession and increase the number of short passing. On the other hand, Leicester City FC thought differently about how they should play football. Their tactic was about counter attacking, which means they were mostly defending for most of the times, but when they gain possession, they quickly passed the ball to the front and scored. Leceister City FC won the league that year, but compared to the other 19 teams in the league, their possession was 42.6%, which was the 3rd lowest in the entire league (Premier League, 2015-2016). This proved that possession does not define earning points (wins and draws).

As shown above, there have been different tactics being the trend in different years of time. Then, what is the current trend of football and what are the factors that make you earn points in the league? In this analysis, which factor affects the most for a football team to win points will be discussed through analyzing multiple data related to soccer.

Methodology

Data

There were a total of three different data sets used in this analysis. Three of the data were sets of different leagues including the English Leauge, EPL, the German League, Bundesliga and the Spanish league, LaLiga. Each data set includes over 150 variables that are related to soccer. These data sets are data from the league of 2019-2020. Then, certain variables that are thought to be the most important factors to earn points were selected from the data sets. The variables that are chosen are

- Total_Pass
- Win
- Loss
- Draw
- GF:Goals Scored
- GA:Goals given
- Possession
- Expenditure

Then, three leagues were combined into a one data set and it is shown in Table 1 in the Appendix ####. This combined data set will be used throughout the analysis and it will be modified in different sections.

Model

When choosing the model of the analysis, two different regressions were used to find the best model. The first regression used is a simple linear regression and the other regression used is a mulitple linear regression.

Simple Linear Regression

In a simple linear regression, a relationship between two variables can be found. The equation of a standard linear regression model is shown below.

$$y = \beta_0 + \beta_1 * x_1 + \epsilon$$

The terms in the equation are: y, β_0, β_1, x_1 and e. Each representing: + y : Dependent Variable + β_0 : Intercept Term + β_1 : Slope Term + x_1 : Independent Variable + ϵ : Residual / Error term

These variables are found using the simple linear regression model and the values are used to find the relationship between the variables and the slope gives whether relation is positive or negative. Simple Linear Regression was used to find the relationship between these three different pairs of variables.

- Points vs Expenditure
- Points vs Total Number of Passes
- Points vs Possession

Multiple Linear Regression

In a multiple linear regression, a relationship between a dependent variable and many independent variables is shown. When MLR is used, using the summary, you can find point estimates of the independent variables and using them, you can predict the outcome of a response variable.

$$y_i = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_i * x_i + \epsilon$$

The terms in the equation are: y, β_0, β_1, x_1 and e. Each representing:

- y_i : Dependent Variable
- β_0 : Intercept Term
- β_1, β_2, β_i : Coefficients / Point Estimates
- x_1, x_2, x_i : Independent Variables
- ϵ : Residual / Error term

Results (3-4 Paragraphs)

Simple Linear Regression Models

The first model that was analyzed using the simple linear regression was finding the relationship between points earned and possession. As it has been mentioned earlier, one of the trends in the 2010s of soccer was having high percentage of possession, so the model was analyzed to check whether that is still correct for current trend of soccer. The Graph 1 in Appendix C shows the graph of Possession vs Points Earned. As it can be seen, there is a positive relationship between the two variables which means as the team gains more possession in games, they earn more points. The correlation of these two varaibles was found, which was 0.7435 and since the value is between -1 and 1 and not close to 0, it means they have high levels of associations between the two variables. Then, the coefficients of this model was found and using the coefficients in Appendix D, the estimated regression line can be written as

$$y = -35.594 + 1.722 * x$$

where x represents the possession and y represents the pointes earned. Using a simple linear regression, it has been found that points earned increase by 1.722 points as the team gains 1% of possession.

A model of finding the relationship between expenditure and points earned was analyzed using the simple linear regression. In graph 2 of the Appendix E, it shows a graph of the relationship between the expenditure of the teams and the points earned. Using the coefficients in Appendix F, we can conclude that the estimated regression line of the model is

$$y = 41.5735 + 0.1232 * x$$

where y is the points earned and x is the expenditure. However, it can be seen from Graph 2 that some teams' expenditure are significantly high compared to the other teams, so these values could affect the slope of the regression line to increase. Hence, another model was analyzed after filtering out 11 teams that had high expenditure. This model is shown in Graph 3 of Appendix G and using the coefficients in Appendix H, we can find that the estimated regression line of the model is

$$y = 43.27156 + 0.07026 * x$$

and we can see that the slope has decreased and from the graph it is visible that the slope is close to being horizontal. This informs us that after excluding teams who spend significantly high amount of money, the points eanred does not change by a lot when the teams spend more. Also, in the Appendix H, it shows that correlation between the two variable is about 0.18 and it is close 0, which means the two values are not correlated. Hence, the 'Expenditure_euro' variable will be excluded from the multiple linear regression that will be performed.

Multiple Linear Regression Models (MLR)

After removing the expenditure variable from the dataset, mulitple linear regression was performed to different models. The variables that will be selected for multiple linear regression are

- Shots
- Total Number of Passes
- Goals Given
- Possession

and these variables were chosen because shots represent values for the attack, number of passes represent values for the midfielders, goals given are the values for the defenders and possession represents the entire team. Although there are many other variables for soccer, these variables were selected to represent the factors that can affect the gameplay.

The first MLR was done with using these four variables to find its effect on the pointes earned and this model can be seen in the Appendix I. By looking at the summary and the p values, we can see that the p-value of possession is significantly high (0.63731). Since this is greater than the alpha, 0.05, this means that the possession variable is not strong enough to suggest and effect exists in the points earned. Hence, the variable has to be excluded from the model.

After removing the variable, MLR was used against to find the relationship between the variables and this model can be found in the Apeendix J. In this model, three variables: Passes, Shots, Goals Against were used and by looking at the p-values, it can be concluded that all of these three variables are highly related to the dependent variable, Point Earned. Additionally, the Multiple R-squared value was 0.8676 and this means that approximately 86.76% of variation in Points earned can be explained by our model. Lastly, using the coefficients of the model given in Appendix J, estimated regression line is found to be

$$y_{PointsEarned} = 54.7447 + 0.0011237 * x_{pass} + 0.0524 * x_{shots} - 0.8505 * x_{GoalsGiven}$$

Discussion

Summary

Throughout this analysis, data of three soccer leagues: English, German and Spanish were used to find what factors affect the teams to earn points. These data were combined and certain variables that were thought to be relevent were chosen. Using this data two types of regression models were performed. In the simple linear regression models, expenditure variable and possession variable was used to find the relationship between the points earned and them. However, in the multiple linear regression model, total number of passes, shots taken and goals given were used to find the relationship between the points earned and them.

Conclusion

In conclusion, the datasets of different soccer leagues were merged to increase the number of data and the accuracy of the analysis. Using this data, we found that possession and points earned have a positive relationship. Meanwhile it was found that expenditure doesn't affect the points earned by a lot after removing certain teams who spend billions of euros. This gave us a fact that current trend of soccer is not always about the money and the actual game play is more important than how much each team spends.

After doing the analysis of multiple linear regression with four variables, we have found that the possession actually does not affect how much the team earns points. From this, it can be concluded that recent trend of soccer has not changed since the 2016 which was the year that Leicester City FC won the league with having low percentage of possession. After removing the possession variable, it was found that the three variables: passes, shots, goals given are the most important factors for teams to earn points in soccer games. The regression line informs us that every pass the team makes, the points earned will increase by 0.00112 and every shot the team takes, the points earned will increase by 0.0524. On the other hand, the points will decrease by 0.8505 as the other teams score on the team. Also, from the Residuals vs Fitted Graph and the Normal Q-Q Graph in the Appendix K, it can be concluded that this multiple linear regression model gives a fairly good explanation of what factors affect the points earned.

From this analysis, it can tell the teams how to play soccer currently. The recent trend of soccer to earn points is, since the possession does not affect the results by much, try to pass the ball fastly and take a shot, but focus on defending. Again, from this, it can be proved that Leceister City FC's style of play is still the trend of soccer in the three leagues. Hence, the teams should focus on playing similar to Leceister City FC.

Weakness & Next Steps

Although linear regressions were used in this analysis, soccer is way more complicated than finding which factor affects the points earned. There could be attendance, team satisfactory, weather, wage and many other factors. However, only certain variables were chosen in models to come up with a conclusion, so that is the weakness of this analysis. For future analysis, more variables can be considered with more number of data sets, which can give an idea of how to run a football team with what kind of tactics to successfully lead a soccer team.

References

- Bairner, R. (2020, May 8). What is tiki-taka? How tactics made famous by Barcelona and Spain work. What is tiki taka? How tactics made famous by Barcelona and Spain work | Goal.com. https://www.goal.com/en-ng/news/what-is-tiki-taka-barcelona-spain-tactics/5f3qumd4uank198jwik1ww8mr.
- Connelly, B. (2020, April 19). How soccer has changed in the past 10 years: From Mourinho's peak to reign of superclubs. ESPN. https://www.espn.com/soccer/english-premier-league/story/4086497/how-soccer-has-changed-in-the-past-10-years-from-mourinhos-peak-to-reign-of-super-clubs.
- Premier League. 2015-2016 Premier League Player Stats. FBref.com. https://fbref.com/en/comps/9/1467/stats/2015-2016-Premier-League-Stats.
- Siregar, C. (2018, December 14). What is Total Football? Famous tactics explained: the clubs, countries & players to use it. https://www.goal.com/en/news/what-is-total-football-famous-tactics-explained-the-clubs/w5yd5dzofn4p17ewmav4atk2b.

Appendix

Appendix A

```
#Modifying the EPL expenditure numbers in the right way of writing because the units were in million
EPL_expenditure_Final <- EPL_expenditure_19_20 %>%
  mutate(Expenditure_euro = Expenditure * 1000000 ) %>%
  mutate(Income_euro = Income * 1000000  ) %>%
  mutate(Balance_euro = Balance * 1000000 ) %>%
  select(Club, Expenditure_euro, Income_euro, Balance_euro, Arrivals, Departures)

#Modifying the Bundesliga expenditure numbers in the right way of writing because the units were in mil
Bundesliga_expenditure_Final <- Bundesliga_expenditure_19_20 %>%
  mutate(Expenditure_euro = Expenditure * 1000000  ) %>%
  mutate(Income_euro = Income * 1000000  ) %>%
  mutate(Balance_euro = Balance * 1000000 ) %>%
  select(Club, Expenditure_euro, Income_euro, Balance_euro, Arrivals, Departures)

#Modifying the LaLiga expenditure numbers in the right way of writing because the units were in million
LaLiga_expenditure_Final <- LaLiga_expenditure_19_20 %>%
  mutate(Expenditure_euro = Expenditure * 1000000  ) %>%
  mutate(Income_euro = Income * 1000000  ) %>%
  mutate(Balance_euro = Balance * 1000000  ) %>%
  select(Club, Expenditure_euro, Income_euro, Balance_euro, Arrivals, Departures)


#Combining the expenditure dataset and general stats expenditure together
EPL1920 <- merge(EPL_19_20, EPL_expenditure_Final, by = "Club")
Bundesliga1920 <- merge(Bundesliga_19_20, Bundesliga_expenditure_Final, by = "Club")
LaLiga1920 <- merge(LaLiga_19_20, LaLiga_expenditure_Final, by = "Club")
```

Appendix B

Table 1

```
##Combining the datasets together with the variables that will be used

EPL1920_tgt <- EPL1920 %>%
  select(Club, Total_Pass, Points, W, GF, GA, Shots, Poss, Expenditure_euro)

Bundesliga1920_tgt <- Bundesliga1920 %>%
  select(Club, Total_Pass, Points, W, GF, GA, Shots, Poss, Expenditure_euro)

LaLiga_tgt <- LaLiga1920 %>%
  select(Club, Total_Pass, Points, W, GF, GA, Shots, Poss, Expenditure_euro)

League_Combined_Prev <- rbind(EPL1920_tgt, Bundesliga1920_tgt)
League_Combined <- rbind(League_Combined_Prev, LaLiga_tgt)

kable(League_Combined)
```

Club	Total_Pass	Points	W	GF	GA	Shots	Poss	Expenditure_euro
Arsenal	16349	56	14	56	48	401	54.0	160400000
Aston Villa	11530	35	9	41	67	453	43.9	159100000
Bournemouth	11943	34	9	40	65	384	43.8	56450000
Brighton	15696	41	9	39	54	456	52.2	74940000
Burnley	9925	54	15	43	50	384	41.4	13850000
Chelsea	20665	66	20	69	54	619	60.7	45000000
Crystal Palace	12142	43	11	31	50	372	44.5	7600000
Everton	13239	49	13	44	56	465	49.2	121000000
Leicester City	17329	62	18	67	41	533	57.6	104300000
Liverpool	20887	99	32	85	33	585	63.4	10400000
Manchester City	24266	81	26	102	35	730	66.9	159520000
Manchester United	17542	66	18	66	36	528	56.2	226780000
Newcastle United	10500	44	11	38	58	397	38.6	72900000
Norwich City	14610	21	5	26	75	409	49.3	8820000
Sheffield United	12052	54	14	39	39	353	43.1	71500000
Southampton	12481	52	15	51	60	497	49.3	58600000
Tottenham	15794	59	16	61	47	439	52.2	148500000
Watford	11207	34	8	36	64	410	42.5	48000000
West Ham	12388	39	10	49	62	414	44.0	120200000
Wolves	14072	59	15	51	40	453	48.3	121800000
Augsburg	8618	36	9	45	63	358	38.1	31500000
Bayern Munich	20739	82	26	100	32	611	65.6	139500000
Dortmund	20781	69	21	84	41	443	61.0	148500000
Dusseldorf	11476	30	6	36	67	424	45.4	10750000
Frankfurt	12408	45	13	59	60	492	51.0	77340000
Freiburg	12120	48	13	48	47	435	47.9	18500000
Hertha BSC	11480	41	11	48	59	359	44.7	110700000
Hoffenheim	14589	52	15	53	53	444	51.9	55850000
Koln	11292	36	10	51	69	413	46.7	18000000
Leverkusen	19160	63	19	61	44	490	63.8	96000000
Mainz05	10302	37	11	44	65	446	43.2	28700000
Monchengladbach	14479	65	20	66	40	476	52.4	40500000
Paderborn07	11642	20	4	37	74	424	45.7	750000
RB Leipzig	16395	66	18	81	37	539	55.5	76500000
Schalke04	12189	39	9	38	58	383	49.0	26000000
Union Berlin	9503	41	12	41	58	389	41.8	7400000
Werder Bremen	12979	31	8	42	69	419	48.8	13950000
Wolfsburg	11716	49	13	48	46	466	48.3	38800000
Alaves	9724	39	10	34	39	299	41.1	10770000
Athletic Bilbao	12589	51	13	40	38	400	48.7	0
Atletico Madrid	13709	70	18	50	27	437	48.5	245300000
Barcelona	24981	82	25	84	38	491	66.9	290000000
Betis	16291	41	10	48	60	461	57.1	100250000
Celta Vigo	14692	37	7	37	49	358	51.9	24600000
Eibar	10623	42	11	38	56	417	46.3	17300000
Espanyol	12032	25	5	25	58	400	47.3	61500000
Getafe	8444	54	14	42	37	397	44.6	21500000
Granada	9933	56	16	50	45	384	43.7	8750000
Leganes	10467	36	8	30	51	421	43.8	16450000
Levante	12463	49	14	45	53	416	48.4	12600000
Mallorca	11932	33	9	39	65	409	44.6	7500000
Osasuna	11102	52	13	46	54	453	47.5	14200000

Club	Total_Pass	Points	W	GF	GA	Shots	Poss	Expenditure_euro
Real Madrid	20329	87	26	70	25	552	59.1	355500000
Real Sociedad	16055	56	16	53	48	418	56.6	21250000
Sevilla	17667	70	19	53	34	475	58.4	177750000
Valencia	14565	53	14	44	53	326	48.9	75000000
Valladolid	11236	42	9	31	43	370	44.3	1400000
Villarreal	16361	60	18	62	49	474	53.2	44800000

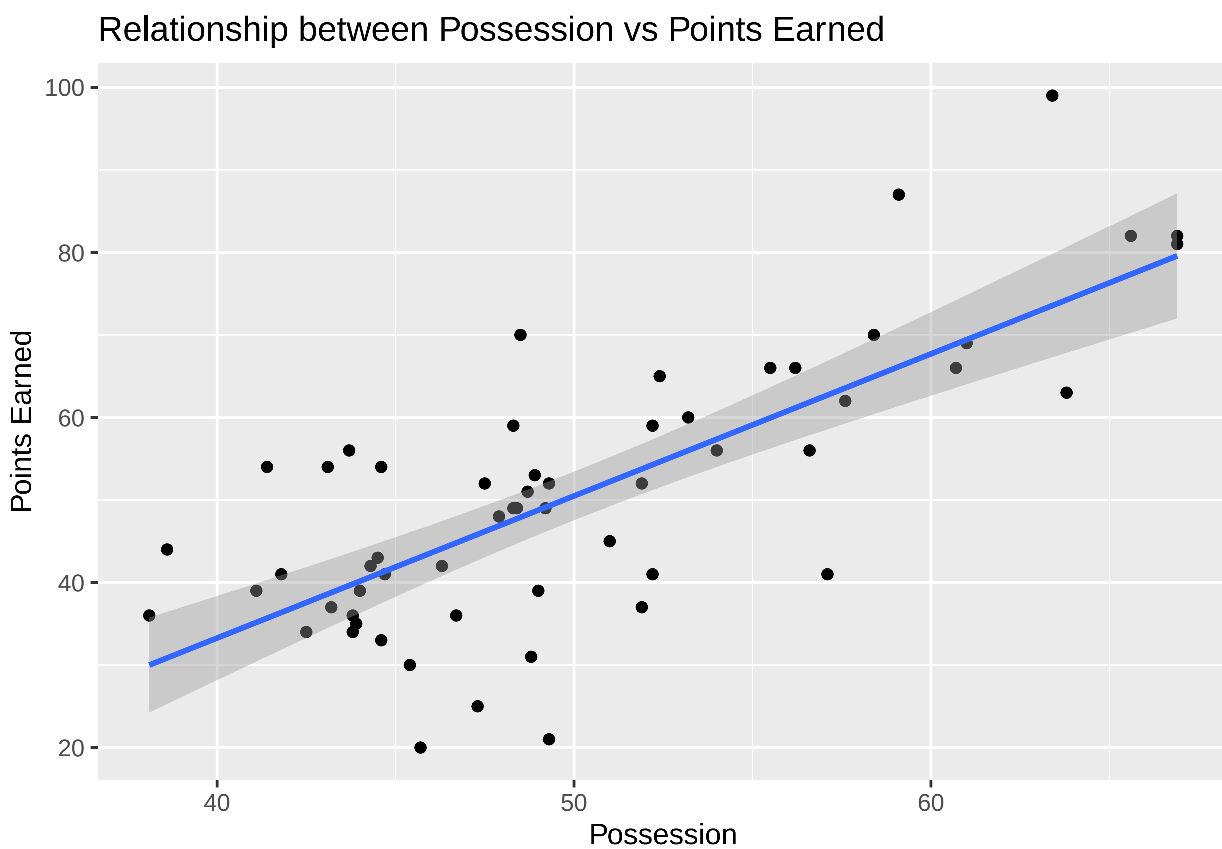
Appendix C

Graph 1

```r
#Simple Linear Regression Comparing Points vs Total Possession


League_simple_poss <- ggplot(data = League_Combined, aes(x = Poss, y = Points )) +
  geom_point() + labs(title = "Relationship between Possession vs Points Earned",
                     y = "Points Earned",
                     x = "Possession")
League_simple_poss + geom_smooth(method = lm)
```

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



Appendix D

```r
#Relevant values to Graph 1, which are correlation and coefficients of Graph 1

coefficients_slr_poss <- lm(Points ~ Poss, data = League_Combined)
coefficients_slr_poss
```

```
##
## Call:
## lm(formula = Points ~ Poss, data = League_Combined)
##
## Coefficients:
## (Intercept)         Poss
##     -35.594        1.722
```

```r
cor(League_Combined$Poss, League_Combined$Points)
```

```
## [1] 0.7435349
```

Appendix E

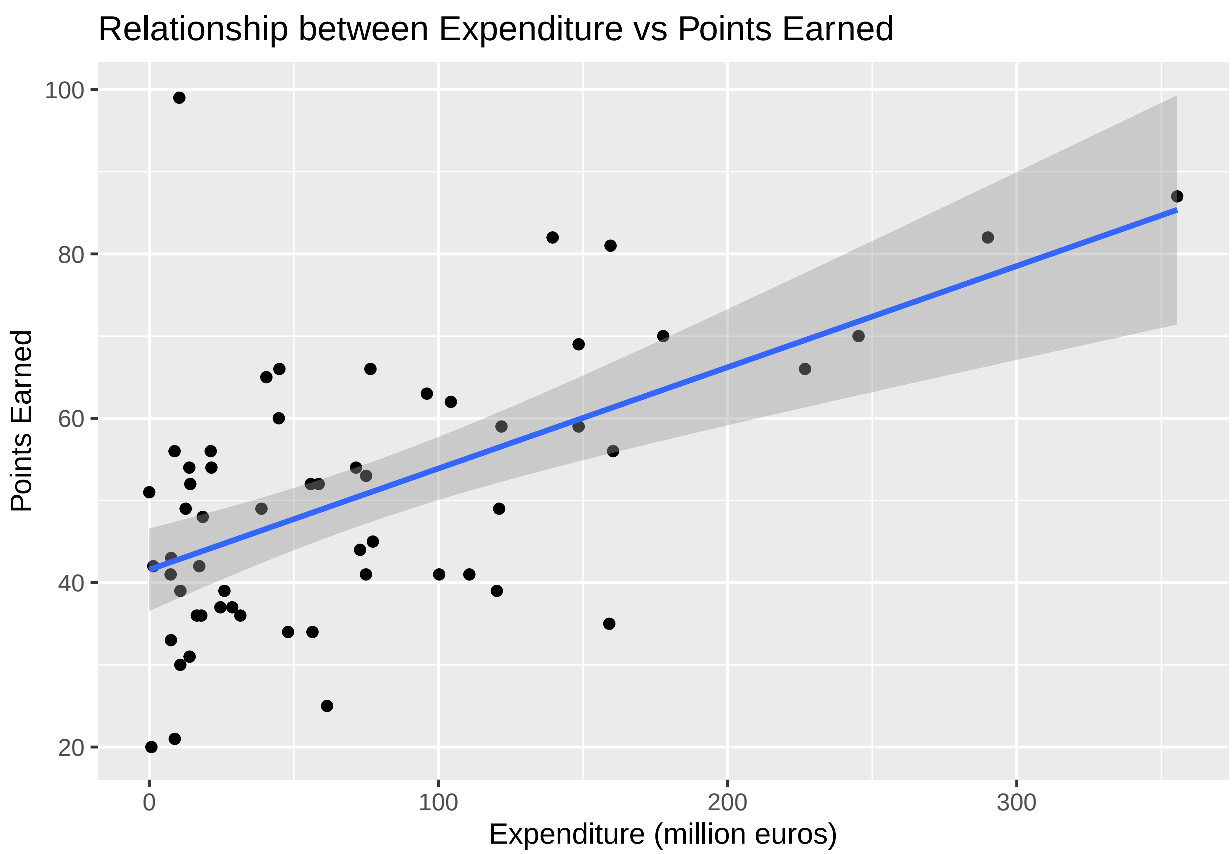
Graph 2

```r
#Simple Linear Regression Comparing Points vs Expenditure

League_Combined_expenditure <- League_Combined %>%
  mutate(expenditure_millions = Expenditure_euro / 1000000)

League_simple_money <- ggplot(data = League_Combined_expenditure, aes(x=expenditure_millions, y=Points))
  geom_point() + labs(title = "Relationship between Expenditure vs Points Earned",
                      y = "Points Earned",
                      x = "Expenditure (million euros)")
League_simple_money + geom_smooth(method = lm)
```

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



Appendix F

```
coefficients_slr_exp <- lm(Points ~ expenditure_millions, data = League_Combined_expenditure)
coefficients_slr_exp
```

```
##
## Call:
## lm(formula = Points ~ expenditure_millions, data = League_Combined_expenditure)
##
## Coefficients:
##          (Intercept)  expenditure_millions
##              41.5735                0.1232
```

Appendix G

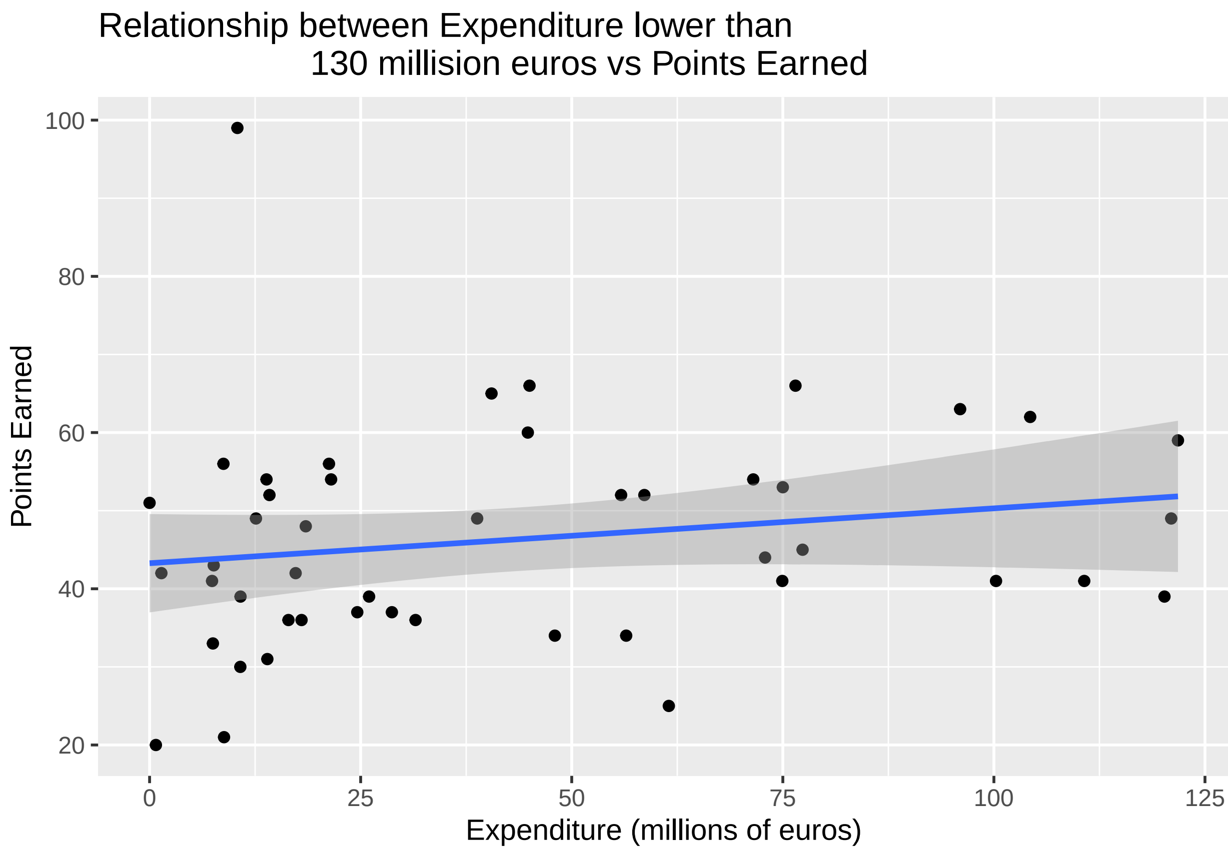
Graph 3

```
League_Combined_expenditure2 <- League_Combined_expenditure %>%
  filter(Expenditure_euro <= 130000000)

League_simple_money2 <- ggplot(data = League_Combined_expenditure2, aes(x=expenditure_millions, y=Points
```

```r
  geom_point() + labs(title = "Relationship between Expenditure lower than
                      130 millision euros vs Points Earned",
                      y = "Points Earned",
                      x = "Expenditure (millions of euros)")
League_simple_money2 + geom_smooth(method = lm)
```

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



Appendix H

```r
coefficients_slr_exp2 <- lm(Points ~ expenditure_millions, data = League_Combined_expenditure2)
coefficients_slr_exp2
```

```
##
## Call:
## lm(formula = Points ~ expenditure_millions, data = League_Combined_expenditure2)
##
## Coefficients:
##          (Intercept)  expenditure_millions
##             43.27156               0.07026
```

```r
cor(League_Combined_expenditure2$Points, League_Combined_expenditure2$expenditure_millions)
```

```
## [1] 0.1857011
```

Appendix I

```r
#Multiple Linear Regression

League_multiple <- lm(Points ~  Total_Pass + Shots + GA + Poss, data = League_Combined)

summary(League_multiple)
```

```
##
## Call:
## lm(formula = Points ~ Total_Pass + Shots + GA + Poss, data = League_Combined)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6893  -3.7077   0.4846   3.3002  18.2386
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.5285361 12.6669993   4.699 1.89e-05 ***
## Total_Pass   0.0014369  0.0007463   1.925  0.05955 .
## Shots        0.0548882  0.0171846   3.194  0.00236 **
## GA          -0.8561294  0.0804237 -10.645 9.03e-15 ***
## Poss        -0.1993151  0.4203193  -0.474  0.63731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.246 on 53 degrees of freedom
## Multiple R-squared:  0.8681, Adjusted R-squared:  0.8582
## F-statistic: 87.23 on 4 and 53 DF,  p-value: < 2.2e-16
```

Appendix J

```r
League_multiple2 <- lm(Points ~  Total_Pass + Shots + GA , data = League_Combined)
summary(League_multiple2)
```

```
##
## Call:
## lm(formula = Points ~ Total_Pass + Shots + GA, data = League_Combined)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9010  -3.7623   0.4829   3.3667  18.1858
##
## Coefficients:
```

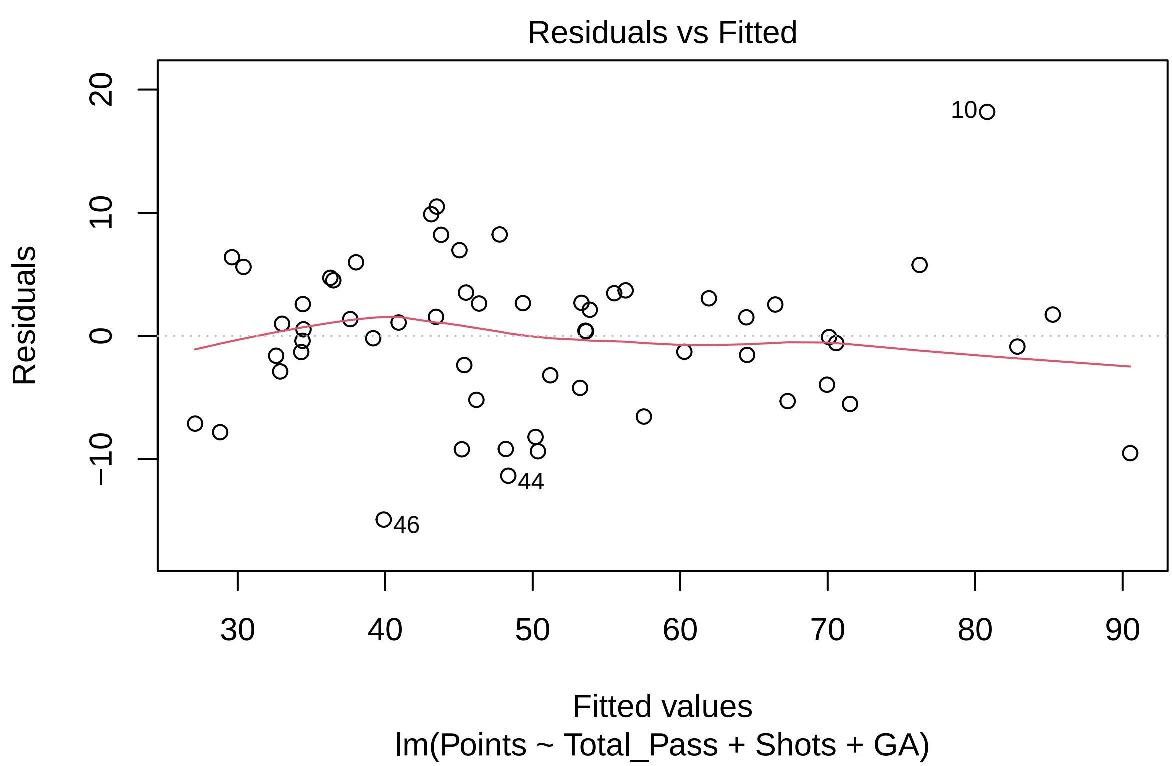
```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.7447074  7.6051469   7.198 1.96e-09 ***
## Total_Pass   0.0011237  0.0003449   3.258  0.00194 **
## Shots        0.0524237  0.0162619   3.224  0.00215 **
## GA          -0.8505771  0.0789937 -10.768 4.67e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.201 on 54 degrees of freedom
## Multiple R-squared:  0.8676, Adjusted R-squared:  0.8602
## F-statistic: 117.9 on 3 and 54 DF,  p-value: < 2.2e-16
```

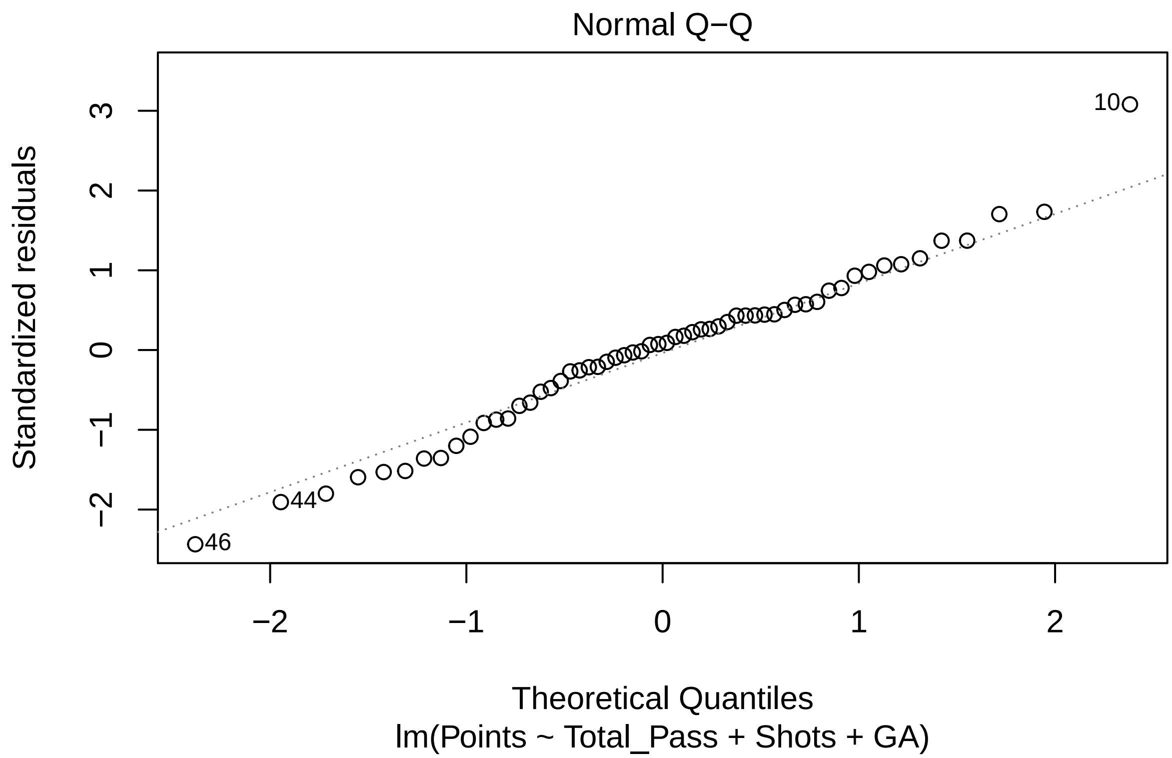
```r
coefficients(League_multiple2)
```

```
##   (Intercept)    Total_Pass         Shots            GA
## 54.744707415  0.001123697  0.052423703 -0.850577100
```

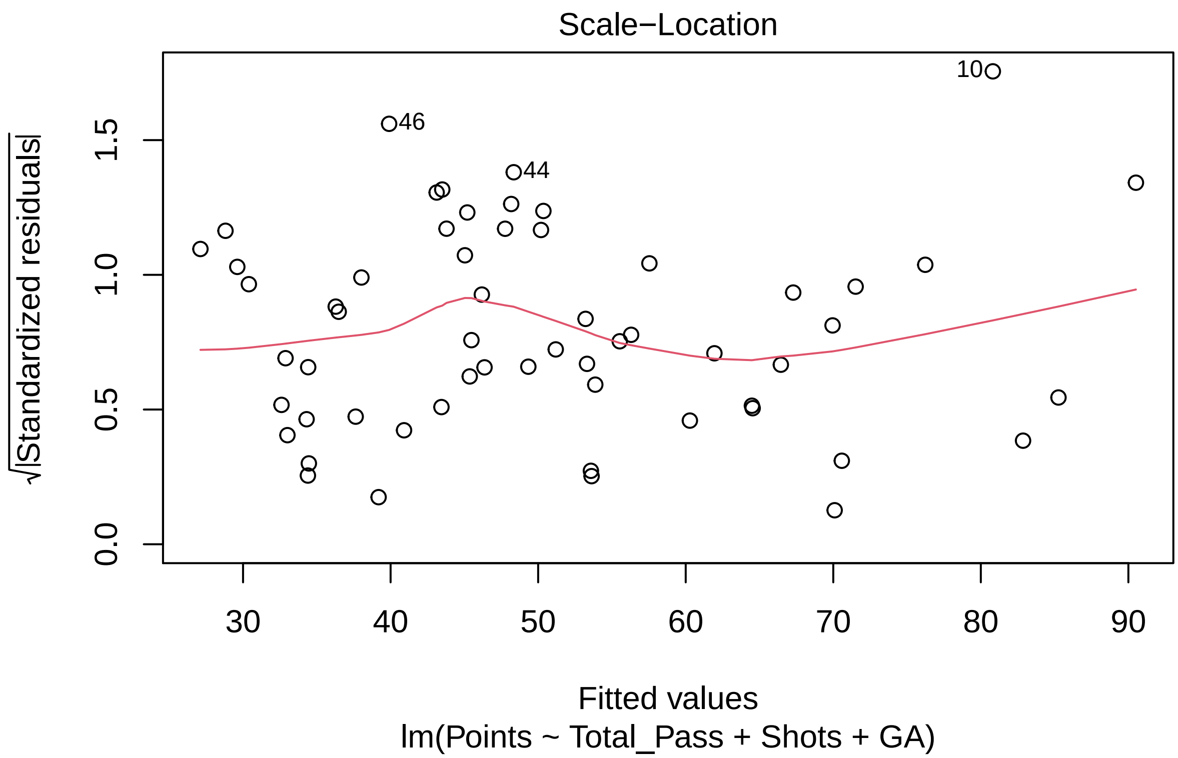
Appendix K

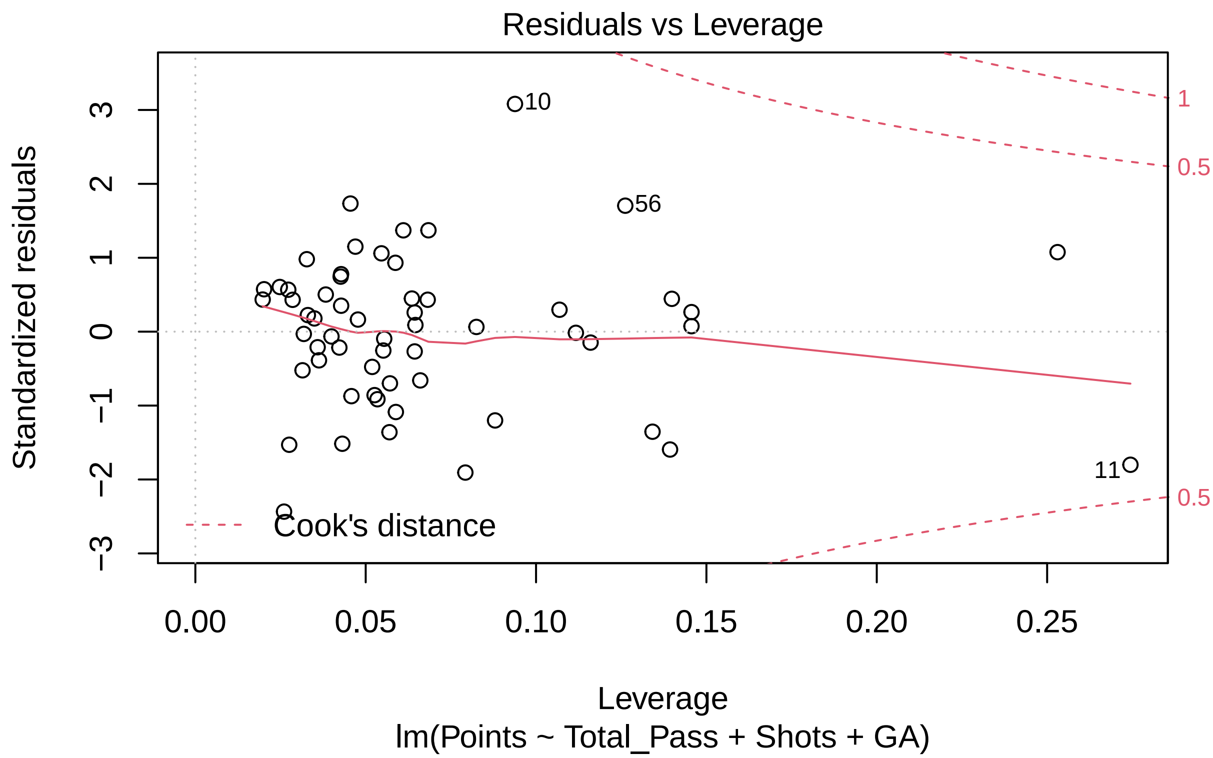
```r
plot(League_multiple2)
```





Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Points ~ Total_Pass + Shots + GA)
10
44
46
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
30
40
50
60
70
80
90
46
44
10
Fitted values
lm(Points ~ Total_Pass + Shots + GA)


Residuals vs Leverage
Standardized residuals
Leverage
lm(Points ~ Total_Pass + Shots + GA)
Cook's distance
10
56
11
1
0.5
0.5
3
2
1
0
−1
−2
−3
0.00
0.05
0.10
0.15
0.20
0.25