Linear Regression to analyze the relationship between points and goal difference in Premier League standings

Zian Chen

Received August 27th, 2020, Accepted May 30th, 2022
First published on the web October 10th, 2020

This paper explains the application of linear regression to analyze the relationship between goals scored, goals allowed, and goal difference with points in the final standings of the English Premier League. The result shows that there is a very strong linear relationship between goal difference and points, as well as a relatively strong linear relationship between goals scored and goals allowed with points. The purpose of this paper is to gain insight into the utility of mathematical applications in analyzing the standings of the Premier League, which can be applied to analyze other soccer leagues. Linear regression is also a useful tool to measure the relative value of attackers and defenders for English Premier League teams under certain assumptions. Given the same capability, a defender should be purchased to achieve a higher standing compared with an attacker.

1 Introduction

In a world full of an enormous amount of data, it is essential to process and analyze them. Linear regression, a mathematical approach that models the relationship between a dependent variable and one or more independent variables[1,2], plays a crucial role in analyzing our daily lives since the method can be employed extensively in practical applications such as in the evaluation of house prices and in identifying differentials between housing areas, as well as personal health conditions and insurance. Linear regression is also significant in the field of machine learning. The linear regression algorithm is a fundamental machine-learning algorithm in view of its relatively simple and widely-known properties[3].

This article mainly focuses on linear regression with one independent variable and one dependent variable, which is called simple linear regression. It concerns two-dimensional sample points in a Cartesian coordinate and predicts the linear relationship between the dependent variable and the independent variable by finding a linear function that best fits the data points[4]. There are several assumptions for the linear regression models, such as constant variance and independence of errors, which means there are no outliers in the data set and the errors of the dependent variables are uncorrelated with each other[1].

A real life example of the English Premier League standings will also be presented to explore the linear relationship between points and goal difference since a club with a larger goal difference often ends the season with higher points, followed by an extension that addresses the question of whether a team should buy a better attacker or defender. The hypothesis is that points and goal difference have a strong linear relationship, and teams in the English Premier League should prioritize the purchase of defenders over attackers since it is widely known that defense wins championships.

2 Mathematical Methods

2.1 Vectors

The first math concept is called vectors, a column of n numbers. If $\mathbf{x}$ is a vector, then $\mathbf{x}$ is in the form of

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

where $x_1, ..., x_n$ belongs to real numbers. n is called the number of components of $\mathbf{x}$, and correspondingly, $\mathbf{x}$ is an n-vector. The length of an n-vector $\mathbf{x}$ is $\|\mathbf{x}\| = \sqrt{x_1^2 + ... + x_n^2}$. The distance between two n-vectors, $\mathbf{x}$ and $\mathbf{y}$ is the norm of the difference $\text{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{y} - \mathbf{x}\|$

2.2 Linear Regression

For each point given, it is feasible to write the point in the form of $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$ $\begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$ ... $\begin{bmatrix} x_n \\ y_n \end{bmatrix}$. Then we can construct two n-vectors, $\mathbf{x}$ and $\mathbf{y}$ respectively, to represent the collections of $x_1, ..., x_n$ and $y_1, ..., y_n$.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}$$

The best fit line can be expressed in the form of $y = \alpha x + \beta$; accordingly, we can construct an n-vector whose values are the y values of this line corresponding to each x.

$$\begin{bmatrix} \alpha x_1 + \beta \\ \alpha x_2 + \beta \\ \dots \\ \alpha x_n + \beta \end{bmatrix}$$

Since the line best fits the data, α and β should be the numbers that make the distance between **y** and the vector of the function $y = \alpha x + \beta$ as small as possible.

Assume that the function J is defined as the average of the sum of the square of the distance from y to $\alpha x + \beta$, so it can be expressed in the following equation.

$$J(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (\alpha x_i + \beta - y_i)^2 \tag{1}$$

The reason why the function needs to be squared is that the accumulative difference will not be compensated by the negative difference if the total difference is positive. Conversely, the positive difference will compensate for the total difference if the sum of the difference is negative. As a result, when the function is squared, the accumulative difference can only increase, which is convenient for finding the best fit line.

According to the function J in equation 1, there are two observations: The first one is that the α and β that best fits the data are the α and β that minimize J. It is because the function is an expression that represents the total distance between the best fit line and the original points; a shorter distance means the line fits the data points better. The second observation is that the function is in a quadratic form. Every quadratic function can be written in the form of $ax^2 + bx + c$. In the function J, the coefficient before x^2, a, is larger than 0 because α^2 is always positive. Therefore, the function $ax^2 + bx + c$ attains its minimum value when $x = -\frac{b}{2a}$.

If β is regarded as a constant and α is the variable, after expanding the function J in the equation 1, the formula can be written in the following form:

$$J = \frac{1}{n}\Big(\sum_{i=1}^{n} x_i^2\Big)\alpha^2 + \frac{2}{n}\Big(\sum_{i=1}^{n} (\beta - y_i)x_i\Big)\alpha + \frac{1}{n}\sum_{i=1}^{n} (\beta - y_i)^2 \tag{2}$$

The second observation indicates that J can attain its minimum value when

$$\alpha = -\frac{\sum_{i=1}^{n} (\beta - y_i)x_i}{\sum_{i=1}^{n} x_i^2} \tag{3}$$

Likewise, If α is regarded as a constant and β is the variable, the equation 1 can be expressed in the following form:

$$J = \beta^2 + \frac{2}{n}\Big(\sum_{i=1}^{n} (\alpha x_i - y_i)\Big)\beta + \frac{1}{n}\sum_{i=1}^{n} (\alpha x_i - y_i)^2 \tag{4}$$

Now, J can reach its lowest point when

$$\beta = -\frac{1}{n}\sum_{i=1}^{n} (a x_i - y_i) \tag{5}$$

After algebraic manipulations of the equations of the minimum α and β, function J can be minimized with the α and β satisfying the following system of equations:

$$\Big(\sum_{i=1}^{n} x_i^2\Big)\alpha + \Big(\sum_{i=1}^{n} x_i\Big)\beta = \sum_{i=1}^{n} x_i y_i \tag{6}$$

$$\Big(\sum_{i=1}^{n} x_i\Big)\alpha + n\beta = \sum_{i=1}^{n} y_i \tag{7}$$

In order to solve the system of linear equations of two unknowns, α and β in this case, the solution formula for equations with two variables can be recalled. Suppose the system of equations is

$$a_{11}\alpha + a_{12}\beta = b_1 \tag{8}$$

$$a_{21}\alpha + a_{22}\beta = b_2 \tag{9}$$

The system of equations 8 and 9 has exactly one solution if and only if

$$\Delta = a_{11}a_{22} - a_{12}a_{21} \neq 0 \tag{10}$$

Then the solution will be

$$\alpha = \frac{a_{22}b_1 - a_{12}b_2}{\Delta} \tag{11}$$

$$\beta = \frac{a_{11}b_2 - a_{21}b_1}{\Delta} \tag{12}$$

This solution in equations 11 and 12 concludes the mathematical methods part. An example of the English Premier League will be presented in the next section.

3 English Premier League

The Premier League is the top level of the English soccer pyramid in which twenty teams compete against each other, playing each other team twice for a total of 38 games. In general, the team with the larger goal difference, calculated as the number of goals scored minus the number of goals conceded, will often win more games. Winning more games will lead to higher points earned because a win game is worth 3 points whereas a draw game is worth 1 point and a lose game is worth 0 points. Therefore, there should be a linear relationship between goal difference and points earned by a team at the end of the season.

In order to explore the relationship, the mathematical methods in section 2 are applied. The standing of the Premier League in the 2018-2019 season is randomly chosen to be the

Clubs	Points	Goal Difference	Goals Scored	Goals Allowed
Manchester City	98	72	95	23
Liverpool	97	67	89	22
Chelsea	72	24	63	39
Tottenham	71	28	67	39
Arsenal	70	22	73	51
Manchester United	66	11	65	54
Wolves	57	1	47	46
Everton	54	8	54	46
Leicester City	52	3	51	48
West Ham	52	-3	52	55
Watford	50	-7	52	59
Crystal Palace	49	-2	51	53
Newcastle	45	-6	42	48
Bournemouth	45	-14	56	70
Burnley	40	-23	45	68
Southampton	39	-20	45	65
Brighton	36	-25	35	60
Cardiff City	34	-35	34	69
Fulham	26	-47	34	81
Huddersfield	16	-54	22	76

Table 1 2018-2019 Season Premier League Standings[5]

example in this section. Table 1 below is created to record the points, goal difference, goals scored and goals allowed, which corresponds to the second to the fifth column respectively.

In this section, since we will explore the impact of the goal difference on the points in the final standings, the independent variable, the x-axis, is goal difference and the dependent variable, the y-axis, is points. Each data point is imported into the plot as Figure 1 shown, as well as the best fit line. After calculation, the coefficient of determination is 0.982, which indicates that there is a extremely strong linear relationship between these two variables.

α and β are calculated, which are 0.64 and 53.45 respectively by using the equation discussed in Section 2. Noticeably, the best fit line printed in the plot has the gradient equal to alpha and y-intercept equal to beta. Since the x-axis is the goal difference and the y-axis is the points of a club at the end of the season, the gradient represents that if a club has one more goal difference, then on average, it will end the season with 0.64 more points. The y-intercept indicates that if a club scores as many goals as it concedes, it will end the season with 53.45 points.

To generalize the conclusion on the linear relationship between points and goal difference, four more years of the standing of the Premier League are analyzed, shown in the appendix of additional data. In the season of 2014-2015, the coefficient

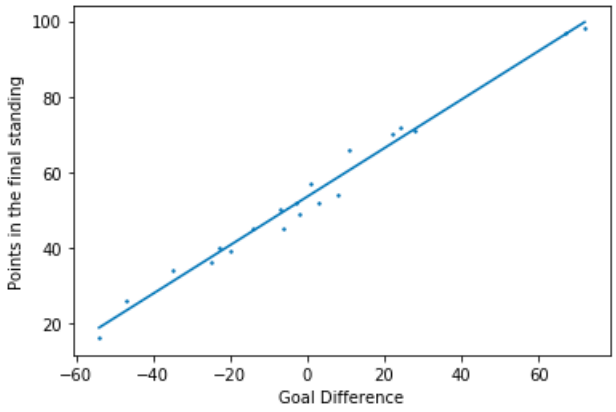

Fig. 1 Goal Difference vs. Points in 2018-2019 Season Premier League

of determination is 0.928. Even though the value of the coefficient of determination is slightly smaller than that in the 2018-2019 season, it is still larger than 0.9, indicating that there is an extremely strong linear relationship between these two variables in 2014-2015 season. The value of α and β are 0.689 and 52.35. Compared with the α and β in 2018-2019 season, one more goal difference would bring slightly more points to the team. Similar to the 2014-2015 season, the

coefficients of determination for the 2015-2016, 2016-2017, 2017-2018 seasons are 0.923, 0.957, and 0.9694. The αs and βs are respectively 0.649 and 51.65, 0.626 and 52.8, and 0.611 and 52.05. All of the coefficients of determination for the five seasons are larger than 0.9, it can be concluded that there is a strong linear relationship between goal difference and points in the standings. The average slope is 0.643 and the average intercept is 52.46, meaning that one more goal difference would bring 0.643 points to the team and a team finishing the season with zero goal difference would end up a season with 52.46 points.

This revelation of the linear relationship between the goal difference and the points inevitably leads to another question for clubs: whether they should spend money on an attacker or spend the same amount of money on a defender, assuming that these two players are identical, with the same capability and efficiency, but playing in different positions. Thus, in order to delve into the question, we can still use linear regression to analyze it quantitatively.

4 Attackers or Defenders?

Similar to the procedure in section 3, the only variable that requires changing is the x-axis variables. Instead of goal difference, it will be substituted by goals scored and goals allowed. Admittedly, some teams which end the season with higher ranking do not necessarily have more goals scored than those teams with fewer points since they can allow fewer goals scored. Conversely, some teams with more points may have more goals allowed because they can score plenty of goals to compensate for the goals allowed. Nevertheless, after plotting each point in the graph, there is still a linear relationship but with a weaker correlation compared with the correlation in section 3. Figure 2 is the plot of goals scored vs. points with the coefficient of determination of 0.943, and Figure 3 is the plot of goals allowed vs. points with the coefficient of determination of 0.850. The slope of the best fit line in Figure 2 is positive because more goals should lead to a higher points whereas the slope of the best fit line in Figure 3 is negative because fewer goals allowed will lead to a higher ranking.

In Figure 2 and 3, the output alpha values are 1.129 and -1.230 respectively, which means one more goal scored can help the team end the season with 1.129 points higher and one fewer goal allowed will enable the team to finish the season with 1.230 points higher. Generally, losing a goal and scoring a goal will be compensated, which means the point difference should be 0 instead of around 0.001, because in the real game, after scoring but losing a goal, two teams go back to the same starting line instead of two teams both losing 0.001 points. When 95 percent confidence intervals are constructed for goals scored and goals allowed, which are from 0.992 and -1.485 to 1.266 and -0.974, the difference of 0.001 is included

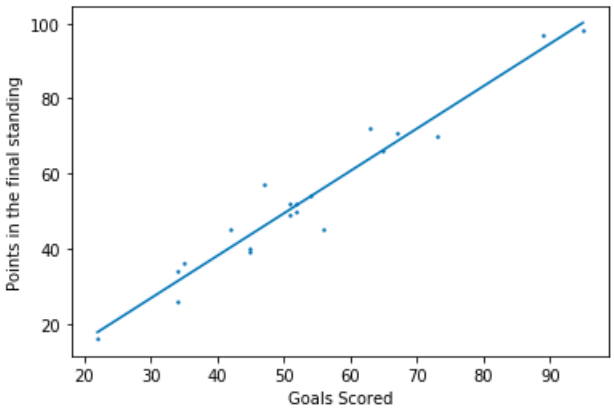

Fig. 2 Goals Scored vs. Points

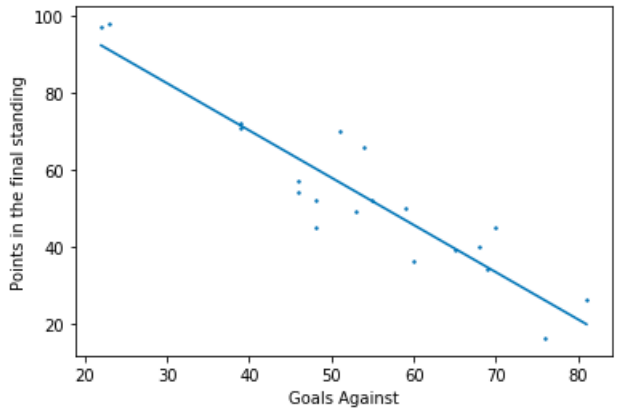

Fig. 3 Goals Allowed vs. Points

in the confidence intervals. Thus, this 0.001 of deviance can be attributed to random errors. This result is reasonable because an attacker or a defender with the same efficiency would have the same effect on the team in essence. When there is a tie situation, if an attacker scores a goal, it would help the team to get 2 more points. When there is a one-goal conceded situation, if an attacker scores a goal, it would help the team to get 1 more point. In contrast, when there is a one-goal leading situation, if a defender intercepts a goal, it would help the team to get 2 more points. When there is a tie situation, if a defender intercepts a goal, it would help the team to get 1 more point. Thus, it would be the same for an efficient attacker or defender in terms of the effect on points. The expectation points for a team with 0 goal difference at the end of the season would be 4/3 because the team can end up having a lose, tie, or win game, with 0, 1, or 3 points respectively. In a win game, the goal difference will be compensated by a lose game. After 38 games in a season, the expectation of total points for a team

would be 50.67, which is very close to our average intercept of 52.46.

Back to the question proposed at the end of the section 3 on whether a club should buy an attacker or a defender, in terms of the quantitative results of the linear regression, a defender should be purchased since a goal allowed is slightly worth more points than a goal scored. In addition, according to the transfer record of the top 10 expensive players in the Premier League shown below, there is only 1.5 defensive player if a midfielder is counted as a half defensive player. Since there is no obvious difference between an attacker and a defender in terms of points brought to the team, purchasing a defender would be a more economic option considering its higher marginal benefit of points per unit of transfer fee.

Player Name	Position	Fee (dollars)
Philippe Coutinho	Left Winger	148.50m
Jack Grealish	Left Winger	129.25m
Eden Hazard	Left Winger	126.50m
Romelu Lukaku	Center Forward	124.30m
Paul Pogba	Central Midfield	115.50m
Gareth Bale	Right Winger	111.10m
Cristiano Ronaldo	Center Forward	103.40m
Harry Maguire	Center Back	95.70m
Jadon Sancho	Left Winger	93.50m
Romelu Lukaku	Center Forward	93.17m

Table 2 transfer record of top 10 expensive players in the Premier League[6]

5 Conclusion

After the plots were generated and analyzed, the final conclusion confirms that the hypotheses are correct since the results prove that there exists a strong linear relationship between goal difference and points, and a defender should be bought to earn a higher standing compared with an attacker in the English Premier League. The sentence ”defense wins championships” is certainly reasonable.

The implication of the research is to better understand the significance of mathematics in real life. Linear regression as an indispensable part of mathematics is a vital and practical tool that has already been widely applied in the fields such as artificial intelligence and business. The paper dives deep into the principles and operations of linear regression, which can further provide readers with insights to comprehend the linear relationships between other variables beyond the English Premier League in soccer. Other leagues such as La Liga, Bundesliga, Serie A, and Ligue 1 can also be applied linear regression to be examined. Additionally, many other real life examples can be quantitatively analyzed by applying the method of linear regression, enabling people to find some particular order in our complex world.

References

1 D. A. Freedman, *Statistical models: theory and practice*, Cambridge University press, 2009.

2 S. Weisberg, *Applied linear regression*, John Wiley & Sons, 2005, vol. 528.

3 T. Kaur, 2018.

4 A. Mehra, *PreMBA analytical methods. Columbia Business School and Columbia University*, 2003.

5 *FBref 2018-2019 Premier League Stats*, https://fbref.com/en/comps/9/1889/2018-2019-Premier-League-Stats.

6 *Premier League - Transfer Records*, https://www.transfermarkt.us/premier-league/transferrekorde/wettbewerb/GB1.

7 *FBREF 2014-2015 premier league stats*, https://fbref.com/en/comps/9/733/2014-2015-Premier-League-Stats.

8 *FBREF 2015-2016 premier league stats*, https://fbref.com/en/comps/9/1467/2015-2016-Premier-League-Stats.

9 *FBref 2016-2017 Premier League stats*, https://fbref.com/en/comps/9/1526/2016-2017-Premier-League-Stats.

10 *FBref 2017-2018 premier league stats*, https://fbref.com/en/comps/9/1631/2017-2018-Premier-League-Stats.

A Additional Data and Graphs

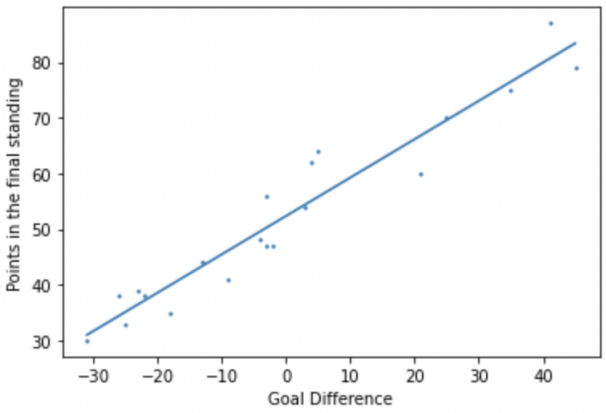

Fig. 4 Goal Difference vs. Points in 2014-2015 Season Premier League

Clubs	Points	Goal Difference	Goals Scored	Goals Allowed
Chelsea	87	41	73	32
Manchester City	79	45	83	38
Arsenal	75	35	71	36
Manchester United	70	25	62	37
Tottenham	64	5	58	53
Liverpool	62	4	52	48
Southampton	60	21	54	33
Swansea City	56	-3	46	49
Stoke City	54	3	48	45
Crystal Palace	48	-4	47	51
Everton	47	-2	48	50
West Ham	47	-3	44	47
West Brom	44	-13	38	51
Leicester City	41	-9	46	55
Newcastle	39	-23	40	63
Sunderland	38	-22	31	53
Aston Villa	38	-26	31	57
Hull City	35	-18	33	51
Burnley	33	-25	28	53
QPR	30	-31	42	73

Table 3 2014-2015 Season Premier League Standings[7]

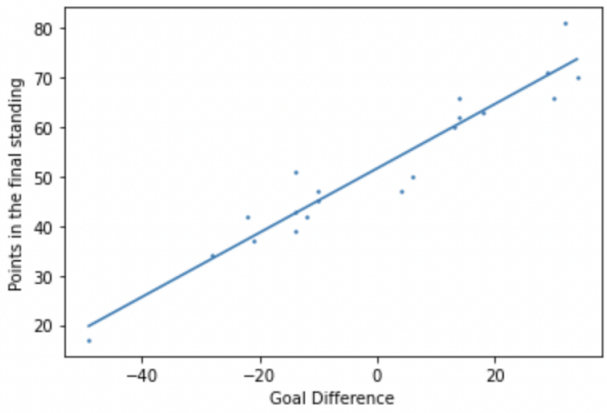

Fig. 5 Goal Difference vs. Points in 2015-2016 Season Premier League

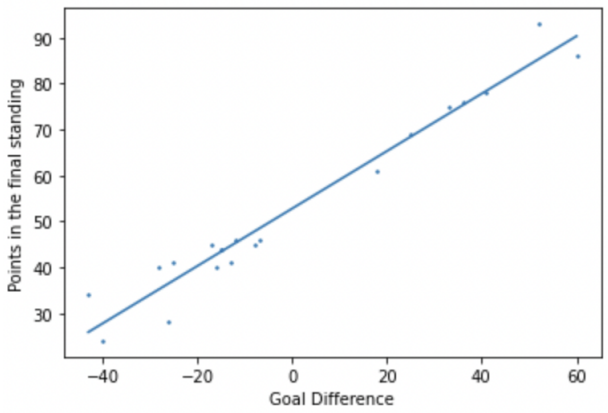

Fig. 6 Goal Difference vs. Points in 2016-2017 Season Premier League

Clubs	Points	Goal Difference	Goals Scored	Goals Allowed
Leicester City	81	32	68	36
Arsenal	71	29	65	36
Tottenham	70	34	69	35
Manchester City	66	30	71	41
Manchester United	66	14	49	35
Southampton	63	18	59	41
West Ham	62	14	65	51
Liverpool	60	13	63	50
Stoke City	51	-14	41	55
Chelsea	50	6	59	53
Everton	47	4	59	55
Swansea City	47	-10	42	52
Watford	45	-10	40	50
West Brom	43	-14	34	48
Crystal Palace	42	-12	39	51
Bournemouth	42	-22	45	67
Sunderland	39	-14	48	62
Newcastle	37	-21	44	65
Norwich	34	-28	39	67
Aston Villa	17	-49	27	76

Table 4 2015-2016 Season Premier League Standings[8]

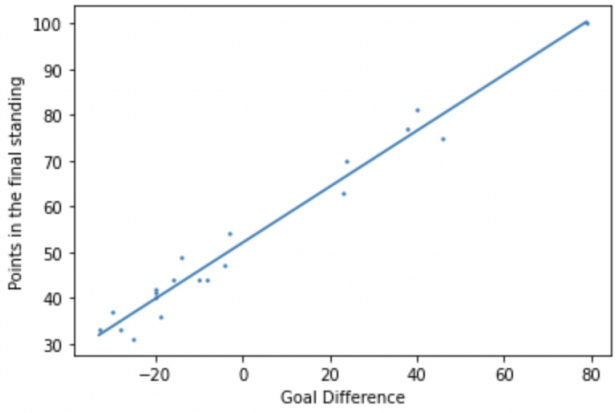

Fig. 7 Goal Difference vs. Points in 2017-2018 Season Premier League

Clubs	Points	Goal Difference	Goals Scored	Goals Allowed
Chelsea	93	52	85	33
Tottenham	86	60	86	26
Manchester City	78	41	80	39
Liverpool	76	36	78	42
Arsenal	75	33	77	44
Manchester United	69	25	54	29
Everton	61	18	62	44
Southampton	46	-7	41	48
Bournemouth	46	-12	55	67
West Brom	45	-8	43	51
West Ham	45	-17	47	64
Leicester City	44	-15	48	63
Stoke City	44	-15	41	56
Crystal Palace	41	-13	50	63
Swansea City	41	-25	45	70
Burnley	40	-16	39	55
Watford	40	-28	40	68
Hull City	34	-43	37	80
Middlesbrough	28	-26	27	53
Sunderland	24	-40	29	69

Table 5 2016-2017 Season Premier League Standings [9]

Clubs	Points	Goal Difference	Goals Scored	Goals Allowed
Manchester City	100	79	106	27
Manchester United	81	40	68	28
Tottenham	77	38	74	36
Liverpool	75	46	84	38
Chelsea	70	24	62	38
Arsenal	63	23	74	51
Burnley	54	-3	36	39
Everton	49	-14	44	58
Leicester City	47	-4	56	60
Newcastle	44	-8	39	47
Crystal Palace	44	-10	45	55
Bournemouth	44	-16	45	61
West Ham	42	-20	48	68
Watford	41	-20	44	64
Brighton	40	-20	34	54
Huddersfield	37	-30	28	58
Southampton	36	-19	37	56
Swansea City	33	-28	28	56
Stoke City	33	-33	35	68
West Brom	31	-25	31	56

Table 6 2017-2018 Season Premier League Standings [10]