

The relationship between team performance and points of soccer teams in La Liga

Ding Yikai

Qingdao No.2 Middle School Post code: 266071

ABSTRACT. *In this paper, the relationship between several factors and team points is estimated. Specifically, the number of wins and goals are proportional to the team points of the season, but pass%, possession% and shots per game have less significance. We research it through simple linear regression, multiple linear regression, t-test and p test. The data are from La Liga, collected from mainly two sources: www.whoscored.com [1] and www.transfermarkt.com [2]. After reading this paper, people will be familiar with the relationship on the large broad. However, this paper has limitations. It cannot be used when emergencies happen. Thus, further research, like influence of covid-19, VAR (visual assistant referee) and weather, need to be discussed in the future.*

KEYWORDS: *Team performance, Data Description*

Introduction

The past few decades have witnessed the rapid development of soccer. Especially with living conditions becoming better and better, people are looking for something for free time entertainment. As a result, watching soccer matches became a wonderful way for people to relax. With soccer competition becoming more and more professional, soccer is not only a way for entertainment, but also for competition. In other words, fans may wonder whether the team will get more points in the season. To predict this, discovering and test the correlation of factors that contribute to the team points is a practical method. Sports data has been a popular topic for statisticians in the recent years. People study various data in many different

ball games and draw conclusions based on their analysis (Yuan He,2012) [3]. To the broadest view, there are many factors that will affect the result of the season, and we will discuss them in sequence in this paper.

We will choose soccer teams in La Liga as the major research subjects. It is well known that La Liga is one of the greatest soccer leagues around the world. Besides, recently, many famous soccer players have played in this league, such as Cristiano Ronaldo (RM), Leo Messi (BAR), Neymar (BAR), etc. Thus, this league has kept the high-quality competitions for a long time and represented the best soccer competitive level.

It is a fact that many factors can contribute to the result of the competition, most of which can be divide into in-field factors and out-field factors. Each factor has its own significance. The variables show off in the figures are representative because of the frequent appearance in team and player ranking. However, to reach our purpose, which is to discuss the effect to the team points, team performance is the most important factor among them.

Team performance has a specific definition, which represent the contribution of soccer players in a match or several matches, like goals, passes, tackles, fouls, etc. To research the importance of each factor, we will use Simple Linear Regression as the major method. To be specific, data will be input into the model. Then, we are able to analyze whether the factor is strongly related to the result of the season.

After analyzing above data, we need to discuss about other events that may influence the performance of the team, including weather, pandemic, the quality of their rivals and the number of kinds of matches they join(league, cup match, or both)etc. Finally, we give the conclusion of the analysis and try to predict the points the teams will get in the next season according to the statistics model we use in the paper.

Through this paper, we can contest the relationship between factors and team points. More goals and wins represent more team points; the pass%, possession% and shots per game are less significant. In addition, this paper will provide soccer fans a new way to predict the team points in the next season.

Data Description

The dataset has 201 rows, each represents the performance of soccer club in La Liga of a specific season. Specifically, the data are from recent 10 seasons, from 2010-2020. The dataset has 13 columns, each represents a factor that may contribute to the result of season.

Data show the competing ability of La Liga, which represent the highest-level soccer match. Besides, the exchange and transfer of players are very common in this league, so it can be a perfect sample data for us.

The majority of data are collected from two websites: the in-field performance data are from www.whoscored.com and www.transfermarkt.com contributes to the complementary data. A glance of data frame is showed in figure 1. Above two websites are the most authoritative ones, so there are few misleading data, which are especially beneficial for our research.

	Team	Goals	Shots per 2	W3	Possession %	Pace%	P	W	D	L	GP	GA	GD	PTS
1	Athletic Bilbao	41	10.8	13	49	76.4	28	13	12	13	41	38	3	51
2	Atletico Madrid	51	11.8	18	45.9	75.8	35	18	16	4	51	27	24	70
3	Barcelona	66	13	25	63.2	88.8	38	25	7	6	56	38	48	82
4	Celta Vigo	37	9.7	7	51.3	80.5	38	7	16	15	37	49	-12	37
5	Deportivo Alaves	34	8.2	10	43.2	70.2	38	10	9	19	34	58	-58	39
6	Eibar	29	11.2	11	46.7	69.2	28	11	9	18	29	56	-17	42
7	Espanyol	27	10.8	5	47.5	74.6	38	5	10	23	27	58	-31	25
8	Getafe	43	10.9	14	43.8	61.9	38	14	12	12	43	37	6	54
9	Granada	22	10.4	16	45.2	71.3	38	16	8	14	22	45	7	56
10	Leganes	30	11.4	8	45	72.2	38	8	12	18	30	51	-21	36
11	Levante	47	11.2	14	48.8	77.9	38	14	7	17	47	53	-8	49
12	Malloca	40	11.1	9	46.2	77.6	28	9	6	23	40	65	-25	23
13	Osasuna	46	12.2	13	47.5	70.5	38	13	13	12	46	54	-5	52
14	Real Betis	45	12.6	10	55.7	83.4	38	10	11	17	45	60	-12	41
15	Real Madrid	70	14.9	28	57.1	86.9	38	28	9	3	70	28	48	87
16	Real Sociedad	58	11.3	16	54.6	81.1	38	16	8	14	58	48	8	56

figure1: the first 17 lines of the points and performance data of teams in La Liga

Variables Selection

P: Plays, the number of the matches that the team plays (In La Liga, there are 20 teams in total, so each team has to play 38 games per season. However, in Bundesliga, there are 18 teams in total, hence each team has to play only 34 games per season.)

W: Wins, the number of matches that the team wins (less than or equal to total number of plays)

D: Draws, the number of matches that the team draws (less than or equal to total number of plays)

L: Loses, the number of matches that the teams lose (less than or equal to total number of plays)

GF: Goal For, the total number of goals that the team makes

GA: Goal Against, the total number of goals that the team loses

GD: Goal Difference, the difference between GF and GA, which can be calculated through the formula GF-GA

Shots pg: shots per game, which represents the number of shots the team makes in one game

Pass%: the percent of successful passes the team makes in one game, which can be calculated by using the formula

Pass% = (the number of successful passes in one game)/(the number of passes in one game)

Possession%: the percent of time that a team control the soccer ball in one game, which can be calculated by the formula

Possession% = (the time that a team control the ball in one game)/(the full time of the game)

Pts: the points of the team in one season (shown in the points table)

GF, GA, GD are chosen to be the variables because of their significance to rank the general ability of a team. P, W, D, L can tell the team points most directly. Also, shots pg, possession%, and pass% are important because they represent the competing style. Thus, none of the factors are unnecessary.

Besides, because of the countless terminations in soccer field, we cannot explain them all. Others will be discussed below in this paper.

Method

In the research, simple linear regression and multiple linear regression are the major methods. It is not difficult to understand the linear regression, which can be

used to predict the expectation, according to the real data. Also, with this method, we can tell the relationship between independent variables (GA, GF.....) and the dependent variable(team points). Simple linear regression can be used when purpose is to detect relation between two variables. When there are more than two variables, multiple linear regression is a better choice. To test whether the conclusion is precise, we select R square to check it. Additionally, in order to further verify the result, we get t value from linear regression model, and then get p value. Based on the above method, we can clearly predict the relationship between variables.

Linear Regression Model Results

The relationships between dependent variables and factors showed by using the single linear regression model below

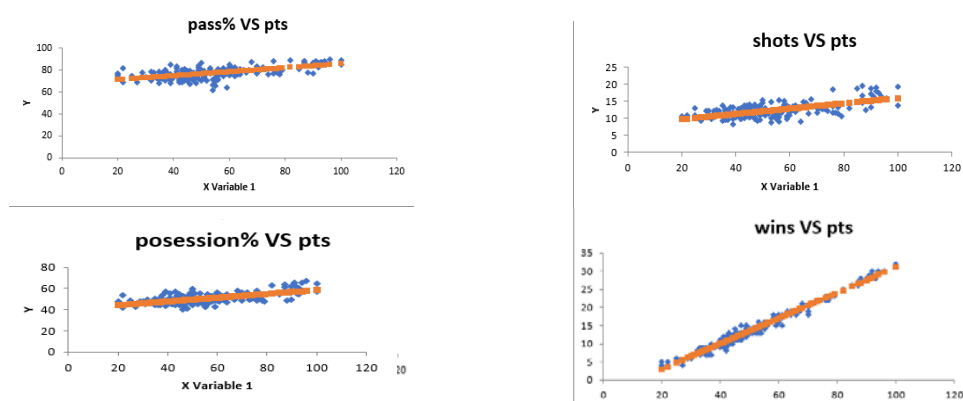


figure 2: team points VS multiple variables

It is apparent that the team points can be influenced by various factors. Take them for examples. When it comes to goals, it is self-evident that, usually, more goals a team gets, more points the team will get. Besides, more wins the team gets, more points the team will get. However, it is another story as for possession%, shots and possession%.

SUMMARY OUTPUT		SUMMARY OUTPUT		SUMMARY OUTPUT	
regression analysis of pass%		regression analysis of possession%		regression analysis of shot	
Multiple R	0.60247864	Multiple R	0.63380106	Multiple R	0.6543376
R Square	0.36298051	R Square	0.40170379	R Square	0.4281577
Adjusted R Square	0.35976324	Adjusted R Square	0.39868209	Adjusted R Square	0.42526961
standard error	4.20734839	standard error	3.78203359	standard error	1.57021892
sample	200	sample	200	sample	200

figure 3: parts of summary output of pass%, possession% and shots per game

The R squares of those simple linear regressions are about 0.4, which means that the relationships between these factors and team points have low importance. The pass% and possession% are high because of the characteristic of La Liga, to control the ball and have fewer passing mistakes. In addition, each team has many shots in one game, since they have aggressive styles and try to make more goals. These conclusions really surprised me. According to the conventions, higher possession% and pass% usually represent formidable ability to control the game and win, like FC Barcelona in 2009-2010 season, in which they got 6 championships (Copa del Rey, La Liga, UEFA Champions League, Supercopa de España, UEFA Super Cup, and FIFA Club World Cup). In that season, FCB got average possession% 63.7% and pass% 87, which are much higher than any team today or in the past.

Besides, the factors are also related with each other, which can be showed via figure4.

SUMMARY OUTPUT		SUMMARY OUTPUT	
regression analysis		regression analysis	
Multiple R	0.904406	Multiple R	0.676507
R Square	0.817951	R Square	0.457662
Adjusted R Square	0.817031	Adjusted R Square	0.454923
Standard Error	2.644228	Standard Error	1.529174
Sample	200	Sample	200

figure 4: the summary output of relationship between predictors

As shown in the figure 4, factors can influence each other. For example, the significance of the relationship between goals and wins is very high ($R^2=0.82$), which means that these two can be considered as one. However, the relationship between shots and possession is not very significant ($R^2=0.45$), so they must be seen as two separate variables and discuss their influence in sequence. In other words, more goals usually represent more wins. In contrast, more shots do not mean higher possession%.

Then, we use multiple linear regression to confirm the result. The data are in figure 5.

SUMMARY OUTPUT						
regression analysis						
Multiple R	0.98017132					
R Square	0.97045989					
Adjusted R Square	0.97790474					
Standard Error	2.59022196					
Sample	200					
variance analysis						
	df	SS	MS	F	Significance F	
regression	5	59124.9605	11824.9921	1762.490961	1.499E-159	
residual	194	1301.59446	6.70924979			
sum	199	60426.555				
	Coefficients	standard error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	14.4057102	3.40222454	4.23420327	3.53881E-05	7.69561326	21.1158072
goals	0.01522644	0.02739713	0.55576773	0.579010023	-0.038808	0.06926091
shots pg	-0.2178083	0.15019097	-1.450209	0.148615047	-0.5140251	0.07840847
wins	2.80867972	0.07036443	39.9161881	1.72E-05	2.66990224	2.9474572
possession%	-0.0688374	0.08418319	-0.8177097	0.414526107	-0.2348692	0.09719435
pass%	0.03710648	0.0722786	0.51338133	0.60826876	-0.1054463	0.17965922

figure 5: summary output of multiple linear regression

As we can see, the multiple R, t-value and p-value all confirm the conclusion that factors like goals and wins are significant. However, pass%, possession% and shots pg are less significant.

Supplement Analysis

As stated previously, so many factors can influence the team points and the in-field performances are just part of them, which leads to the above discussion-----out-field factors. Such contributors can also have significant influence on the season result, because they are always unpredictable and mysterious. Injuries, weathers, transfer news, commercial activities.....those factors are especially important.

As everyone knows, it is harmful and terrible for a soccer player to get injuries. When they get injuries, players have to be cured and cannot appear in the game, which may have negative influence to the result. The injury records of Ousmane Dembélé is shown in the above figure.









Season ↑	Injury ↓	from ↑	until ↑	Days ↑	Games missed ↑
19/20	Hamstring Injury	Feb 4, 2020	Aug 13, 2020	191 days	19 
19/20	Torn muscle bundle	Nov 28, 2019	Feb 2, 2020	66 days	15 
19/20	Muscle Injury	Sep 27, 2019	Sep 30, 2019	3 days	1 
19/20	Hamstring Injury	Aug 19, 2019	Sep 22, 2019	34 days	5 
18/19	Hamstring Injury	May 5, 2019	Jun 16, 2019	42 days	4 
18/19	Torn muscle bundle	Mar 14, 2019	Apr 9, 2019	26 days	4 
18/19	Ankle Injury	Jan 21, 2019	Feb 8, 2019	18 days	5 
17/18	Torn muscle bundle	Jan 15, 2018	Feb 10, 2018	26 days	7 
17/18	Hamstring Injury	Sep 17, 2017	Jan 1, 2018	106 days	20 

figure 6: the injury records of Ousmane Dembélé

As we can see in figure 6, since he joined FC Barcelona (Aug 25, 2017), he has had 9 injury records and miss 80 games. Figure 6 shows the team statistics of Barcelona in recent 5 seasons.

season	goals	league points
15-16	112	91
16-17	116	90
17-18	99	93
18-19	90	87
19-20	86	82

figure 7: the records of goals and league points of Barcelona in recent 5 seasons

It is apparent that since 2017-2018 season (Ousmane Dembélé joined Barcelona), the team statistics changed surprisingly. Barcelona lacks a LW(Light Winger) for the long-time absence of Ousmane Dembélé. In that situation, the forwards hardly focus on attacking rivals' defense and have less opportunities to make goals. As a result, the team points have gone down for many years

Limitation

It is true that the model can be used to predict majority of seasons, since it is conducted based on the large quantity of data from past decade. However, there are some situations that the model is not practical. Take the 2019-2020 season for example. Not only in Spain, every league around the world has been affected negatively. For example, the most popular 5 leagues (Premier League, Ligue 1, La Liga, Bundesliga, and Serie A) and UEFA Champions League have stopped for a long time to avoid infection and protect players' health. Besides, in Chinese Super League, after long-period intermission, league was separated into two independent ones, Group A and Group B. In the first stage, each group has 8 teams, and the top 4 of each group can enter the next stage to fight for the champion. Other team have to compete to avoid degradation. This system is pretty different with the previous one, and affect the team points accordingly. Thus, when it comes to seasons with emergencies, we should be careful whether the linear regression model is precise and efficient.

Future Development

The linear regression model can be used to predict team points of future seasons or analyze the points of team in other leagues (Serie A, Premier League.....).

Further, more factors can be added into the model. For example, recently, the VAR (Video Assistant Referee) was introduced into the competition to help referee judge the game. Some factors are recorded: fouls, goals, penalties, etc. There was a decrease in the number of offsides, fouls and yellow cards after the implementation of the VAR. Meanwhile, there was an increase in the number of minutes added to the playing time in the first half and the full game, but not in the second half. These findings may help coaches, and managers to better understand the effects of the VAR system on professional soccer and to identify strategies to improve refereeing during matches (Lago-Peñas Carlos, Rey Ezequiel and Kalén Anton, 2019) [4]. However, the VAR also has disadvantages. Referees sometimes ignore the situations to check VAR (even if it is a foul, it does not work if referee do not stop the match and check it.). The influence of VAR is not clear now, but majority people believe it

has more benefits than drawbacks. To make the conclusion more precise, VAR can be seen as a factor of this model and be discussed in the future research.

Discussion

Based on the above simple linear regression model, we can definitely conclude that goals and wins can contribute to the team points positively. More goals the team get, more points the team will get in the end of the season; more wins the team get, more points the team will get. However, other factors, such as Pass%, Possession% and Shots pg will not affect the result largely. In other words, above three factors are not significant. There is only little difference whether a team has more shots per game or less shots per game. The situation works for other two factors.

Multiple linear regression is used to further test the result of simple linear regression. Besides, the multiple t-value and p-value strongly verify the conclusion of above paragraph.

Other factors, such as injuries and emergencies like covid-19 also have importance. Injuries can prevent players from joining the game, and influence team points negatively accordingly. The intermission caused by covid-19 make teams and players rest for a long period. However, the influence to team points needs further research.

Reference

- [1] WHOSCORED website, www.whoscored.com
- [2] TRANSFERMARKT website, www.transfermarkt.com
- [3] Yuan He, Advisor: David Aldous, Predicting Market Value of Soccer Players Using Linear Modeling Techniques, 2012
- [4] Lago-Peñas Carlos, Rey Ezequiel and Kalén Anton, How does Video Assistant Referee (VAR) modify the game in elite soccer? , 2019