The Frontiers of Society, Science and Technology
ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

Multiple regression model for predicting soccer player value in English Primer League

Shuangxian Li

Northern Illinois University 60115 America

ABSTRACT: *In this paper, market values of the football players in all teams of English Primer League in 2018-2019 season are estimated using multiple linear regression by including the physical, performance factors and club rank in 2018-2019 season. Players from English Primer League are examined, and by applying VIF test for no collinearity and Durbin–Watson test for no autocorrelation, and the most and the least affecting factors are explained in detail.*

KEYWORDS: *Player value; football club; correlation analysis; multiple linear regression; regression coefficient value; significant level.*

Introduction

Nowadays, soccer association as one of the most popular sports all over the world attracts a lot of fans. Because of its popularity, many professional soccer associations obtain huge incomes. Thus, soccer association is not only a soccer club with fans, but also regarded as a company which including various management of income. These companies always make important decisions about which soccer player they want employ [1].

It has proved that transfers of players make a significant impact on soccer clubs and market value of players provides budget for these transfers [2]. Therefore, player value (market value of a player) is of importance to sales and profit for the companies, which pay attention to analysis and predict player value. So, what kind of factors will influence the player value dramatically?

Obviously, physical condition of players always plays a vital role in studying player value. In football, like many other sports, the age, height, position is related to on-site performance. In the National Football League (NFL), the excellent physical condition is related to the increase in game time and wages [3].

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

Recently, more and more researchers seek for useful factors to analyze player value, Herm think that common indicators of market value including two categories: player characteristics, player performance [4].

Therefore, in this project, we will collect as well as investigate data of player physical condition(including height, age and so on), player performance (including goals, assists and so on) and club rank in all teams of English Prime League competition from 2018 to 2019. Then, analyzing the correlation relationship between the player value and variables of these three categories and predict the player value under the method of multiple regression analysis.

Data used in this project are from transfermarkt.com, English Premier League.com and transferkt.com.

Besides, the result we get is that age, average playing time, average appearance time and club rank make a tremendous effect on market value of soccer players.

This article can present the analysis progress of multiple regression, prove the relationship of player value and player physical condition, player performance and club rank. Furthermore, it can provide reference for soccer clubs to choose suitable players and give reasonable revenue during the transfer.

Method

1. Analysis

This study uses multiple linear regression analysis under quantitative research methodology. Because the multiple regression model not only can help us make predictions about the data, but also can help us to identify the variables that have a significant effect on the dependent variable, it is so suitable as well as reasonable for us to use here. Besides, in regression analysis, if there are two or more independent variables, it is called multiple regression [5]. In fact, a phenomenon is often associated with multiple factors, and it is more effective and realistic to have the optimal combination of multiple independent variables working together to predict or estimate the dependent variable than to use only one independent variable for prediction or estimation. Therefore, multiple linear regression is more practical than one-dimensional linear regression here.

In this project, firstly, we choose 15 variables (including age, height, weight, position, average playing time, average appearances time, average goals, average assists, average clean sheets, penalty goals, penalty misses, red cards, yellow cards, player rank, club rank) as dependent variables. According to recently research, variables such as age, height, position, playing time, goals, assists, cards and other player’s personal

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

information, performance are analyzed and these variables have been demonstrated be the most common indicators of market value [6]. At the same time, club rank is proved that the value of the soccer club always affects the income players want during the transfer season [7]. By selecting different variables, a multiple linear regression model with player value as the dependent variable and player attribute set as the independent variable is finally developed.

The 15 independent variables are all factors that can possibly affect the market value of a soccer players. These 15 predictors can be grouped into three categories:

Dependent variables

Player value: Monetary value of players. Since the value of this variable is too large compared to other variables, we use logarithm of the variables.

Independent variables

(1) Personal data of the player (4 variables)

Age: The age of the players. The age variable was treated as a dummy variable. Players less than 25 years of age were marked as 1; players between 25 and 30 years of age were marked as 2; and players older than or equal to 30 years of age were marked as 3.

Height: The height of the players. The height variable was treated as a dummy variable. Players with a height between 160 cm and 170 cm were marked as 1; players with a height between 170 cm and 180cm were marked as 2; players with a height between 180 cm and 190 cm were marked as 3; and players with a height greater than 190 cm were marked as 4.

Weight: The weight of the players. The weight variable was treated as a dummy variable. Players weighing between 50 kg and 70 kg were labeled as 1; players weighing between 70 kg and 80 kg were labeled as 2; players weighing between 80 kg and 90 kg were labeled as 3; and players weighing more than 90 kg were labeled as 4.

Position: The player's position in the match. The position variable was treated as a dummy variable. Midfielder were labeled as 1; Forward were labeled as 2; Goalkeeper were labeled as 3; Defender were labeled as 4，

(2) Performance data of the player (10 variables)

Average playing time (APT1): Average playing time of players at home and away during one match. Since the value of this variable is too large compared to other variables, we use logarithm of the variable.

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

Average appearances time (APT2): Average number of home and away games played by players.

Average goals (AG): Average number of goals scored by players at home and away.

Average assists (AS): Average number of assists for players at home and away.

Average clean sheets (ACS): Average number of clean sheets for players athames and away.

Penalty goals (PG): Number of penalty kicks taken by players.

Penalty misses (PM): Number of penalty kicks lost by players.

Red cards (RC): Number of red cards awarded to players.

Yellow cards (YC): Number of yellow cards awarded to players.

Player rank (PR): Rank in club top scorer.

3.The ranking of the player's club (1 variable)

Club rank (CR): Clubs' league standings.

Secondly, before building the model, we show some descriptive statistic of these data in order to know about the basic character of them.

Thirdly, Correlation analysis is used to examine the relationship between quantitative data, whether there is a relationship, and how close the relationship is. It can be as preparation before multiple linear regression analysis. It always be measured by Pearson's coefficients (r). The larger the absolute value of Pearson's coefficient indicates the degree of correlation between the independent variable and the dependent variable, while the sign of the value indicates positive or negative.

Finally, the most important step of analysis progress is to build the predict model by multiple linear regression analysis. VIF, R-squared, F-test and DW test can assist the regression. VIF is a collinearity index and more than 5 indicates that there is a collinearity problem. R^2 is coefficient of determination. F test shows that at least one independent variable in the model has a significant influence on dependent variable or shows the significance of the whole model. When analyzing it, we focus on the P value. D-W value is Durbin-Watson test and is a test method of autocorrelation. If the D-W value is around 2 (between 1.7 and 2.3), it means that there is no autocorrelation, and the model is well constructed. Through VIF and D-W test these we can test the autocorrelation and multicollinearity in order to make sure the model is convinced.

2. Hypothesis

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

Player personal information, player performance and club rank can make a significant impact on player value.

3. Data

Because English Primer League (EFL) as a historical European league compared with other leagues, it provides large amount of convincing matches results, player information we decide select data from EFL to analyze. The data selected for this study is the 2018-2019 Premier League players, matches and team data from English Primer League.com, transfermarkt.com and footystats.org. These websites not only contain all available information on soccer sports activity, teams and player information, but also have common rather than special or abnormal data. Therefore, these data can represent the statistics results of majority of soccer data all over the world.

We selected 15 independent variables and one dependent variable from players, matches and team data of English Primer League from 2018 to 2019. The independent variables are divided into three categories: physical data of the players, performance data of the players and the ranking of the players' club. The dependent variable is the players' value. After importing the raw data into Python for data cleaning we get a dataset of 324 rows and 17 columns. As shown in Figure 1.

	Player	Age	Height	Weight	Position	Average playing time	Average appearances time	Average goals	Average assists	Average clean sheets	Penalty goals	Penalty misses	Red cards	Yellow cards	Player rank	Club rank	Player value
0	Aaron Cresswell	3	2	1.0	4	6.677713	9.0	0.0	0.5	1.5	0	0	0	1	20	10	15.581952
1	Andriy Yarmolenko	3	3	3.0	2	5.435903	2.5	1.0	0.0	0.5	0	0	0	2	7	10	15.961442
2	Andy Carroll	3	4	2.0	2	5.424950	1.5	0.0	0.0	1.0	0	0	0	2	17	10	14.873301
3	Angelo Ogbonna	3	4	3.0	4	6.840547	10.0	0.5	0.5	1.5	0	0	0	2	13	10	15.581952
4	Arthur Masuaku	2	2	2.0	4	6.777647	9.5	0.0	0.5	3.0	0	0	0	2	25	10	15.581952
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
329	Matt Doherty	2	3	3.0	4	7.361058	17.5	2.0	2.5	4.5	0	0	0	5	5	7	16.482739
330	Morgan Gibbs-White	1	2	NaN	1	5.779199	2.5	0.0	0.5	2.5	0	0	0	1	16	7	15.684231
331	Romain Saiss	3	4	3.0	1	6.349139	6.0	1.0	0.0	2.0	0	0	0	7	7	7	15.789592
332	Ryan Bennett	3	3	2.0	4	7.333023	17.0	0.5	0.0	4.5	0	0	0	12	10	7	15.096444
333	Willy Boly	2	4	4.0	4	7.368024	18.0	2.0	0.0	4.0	0	0	1	3	4	7	16.384299

324 rows × 17 columns

Figure 1

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

The empirical model for estimation is developed as follows:

$$DV = \beta + \beta_1 Age + \beta_2 Height + \beta_3 Weight + \beta_4 Position + \beta_5 APT1 + \beta_6 APT2 + \beta_7 AG + \beta_8 AS + \beta_9 ACS + \beta_{10} PG + \beta_{11} PM + \beta_{12} RC + \beta_{13} YC + \beta_{14} PR + \beta_{15} CR + \varepsilon$$

Where e is the error term, DV is the dependent variable and the remaining variables are described below.

Result

1. Descriptive statistic

Table 1 below shows that total number, mean, stand deviation, minimum, quartile and maximum of every variable.

Variable	count	mean	std	min	25%	50%	75%	max
Player value	323	15.96	1.06	13.49	15.28	15.96	16.65	18.56
Age	323	2.10	0.70	1.00	2.00	2.00	3.00	3.00
Height	323	2.80	0.72	1.00	2.00	3.00	3.00	4.00
Weight	323	2.20	0.79	1.00	2.00	2.00	3.00	4.00
Position	323	2.36	1.32	1.00	1.00	2.00	4.00	4.00
APT1	323	6.34	1.15	0.00	5.99	6.71	7.13	7.44
APT2	323	9.10	5.79	0.00	4.00	9.00	14.50	19.00

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

AG	323	1.14	1.88	0.00	0.00	0.50	1.50	11.00
AS	323	0.87	1.28	0.00	0.00	0.50	1.00	7.50
AC	323	3.13	2.32	0.00	1.50	2.50	4.50	10.50
PG	323	0.18	0.76	0.00	0.00	0.00	0.00	7.00
PM	323	0.05	0.30	0.00	0.00	0.00	0.00	3.00
RC	323	0.10	0.30	0.00	0.00	0.00	0.00	1.00
YC	323	2.78	2.69	0.00	1.00	2.00	4.00	13.00
PR	323	12.58	7.42	1.00	6.00	12.00	18.50	29.00
CR	323	10.18	5.79	1.00	5.00	10.00	15.00	20.00

2. Correlation analysis

The results of the correlation analysis between each independent variable and the dependent variable are shown in Table 2. The values in the table are Pearson's coefficients(r).

	Player value
Age	-0.349**
Height	-0.01
Weight	-0.053
Position	-0.052
APT1	0.389**

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

APT2	0.403**
AG	0.495**
AS	0.416**
AC	0.577**
PG	0.205**
PM	0.112*
RC	0.045
YC	0.105
PR	-0.358**
CR	-0.509**
* p<0.05 ** p<0.01	

Table2

Note: The larger the absolute value of Pearson's coefficients indicates the degree of correlation between the independent variable and the dependent variable, while the sign of the value indicates positive or negative.

As can be seen in table 2, we can use the correlation analysis to study player value and 15 variables (including age, height, weight, position, AT, APT, AG and so on) respectively.

The value of the correlation coefficient (r) between player value and age is -0.349 and P-value shows significance at the 0.01 level, thus indicating a significant negative correlation between player value and age. Besides, player value is also negatively correlated with PR and CR. At the same time, there is significant positive correlation between player value and six variables (APT1, APT2, AG, AS, ACS, PG) separately. However, height, position, RC or YC do not make any impact on player value because of the larger p-value.

Therefore, we can make a simple and brief judgment about the relationship between player value and 15 independent variables here through the correlation analysis. Then use regression analysis below to further verify the accurate relationship between them.

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

3. Multiple linear regression

The results of the multiple linear regression reintegrate from results of SPSS shown in Table 3.

	Beta	t	P	VIF
Constant	15.746	45.725	0.000**	-
Age	-0.626	-12.319	0.000**	1.092
Height	0.102	1.664	0.097	1.692
Weight	1.117	1.117	0.265	1.64
Position	-0.472	-0.472	0.637	1.29
APT1	0.214	3.887	0.000**	3.463
APT2	0.025	1.921	0.056	4.732
AG	4.915	4.915	0.000**	3.229
AS	0.799	0.799	0.425	1.803
AC	1.754	1.754	0.08	3.076
PG	1.427	-1.427	0.155	2.152
PM	0.385	0.385	0.7	1.54
RC	-0.084	-0.084	0.933	1.162
YC	-0.921	-0.921	0.358	1.757
PR	0.361	0.361	0.718	2.206
CR	-9.927	-9.927	0.000**	1.647
R^2	0.684		F	44.237

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

Adjust R^2	0.668		P	0
DW	1.821			
* p<0.05 ** p<0.01				

Table 3

Note: P- value is obtained by t test. P value less than 0.05 means that the corresponding X has a significant influence on the dependent variable.

we use these 15 predictors as independent variable, and player value is used to as dependent variable in the linear regression analysis. We can see that R-squared is equal to 0.684, which means that the regression equation can explain the 68.4% variation in player value.

The F-test of the model found that the model passed the F-test (F=44.237, p=0.000<0.05), which means that at least one of these 15 independent variables that will influence Player value. Therefore, we can get the predicted model formula below:

In addition, the test for the model's multicollinearity found that all the VIF values in the model are less than 5, so there is no collinearity problem, which indicates that 15 explanatory variables do not have the linear correlation with each other.

The D-W value is near 2, thus indicating that the model does not have autocorrelation, there is no correlation between the sample data. Therefore, the model using multiple regression analysis is good.

The final specific analysis shows that:

Age has a regression coefficient value of -0.626 (t= -12.319, p=0.000<0.01), implying that Age will have a significant negative impact relationship on player value. In addition, the data above shows that APT1, AG also affect player value negatively. Club rank has a regression coefficient value of -0.075 (t=-9.927, p=0.000<0.01), implying that club rank will have a significant negative impact relationship on Player value. At the same time, according to insignificance of the p values, the remaining 7 predictors (height, weight, position, APT2, AS, ACS, PG, PM, RC, YC, PR, CR) will not affect the player value.

From the results and information above, we know that market value can be influenced by many factors. Many predictors were extremely significant. It is apparent

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

that the older you are, the less player value you get; the more you appear in games, the more you score. It is also obvious that players score more goals, the companies may give higher transfer fees for them. However, many of the performance data from last year failed to be significant. Moreover, although value of players was largely affected by their position in many football teams [8], the position factor here has a p-value bigger than 0.10, which means that no impact on the player value of EFL.

4.Discussion

According to the multiple regression method, APT1, AG will have a significant positive effect on Player value which is highly consistent with the hypothesis. Age, CR will have a significant negative effect on player value. However, height, weight, position, APT2, AS, ACS, PG, PS, RC, YC, PR does not have an impact on Player value, which contradicted our assumption. Therefore, when we would like to study the relationship further, we can delete the 11 variables, only remain the other 4 independent variable.

Furthermore, there are also some limitations to this study. Firstly, we only used data from the 2018-2019 Premier League and the results of the study can only illustrate some of the factors that influence the value of players in the 2018-2019 Premier League. Besides, the explanatory power of the model is only 68.4%. The reason for this could be due to the small sample size or the choice of independent variables that do not explain the player value well. So we should delete these useless variables if we want to get the more ideal results of predicting research.

Although there are some limitations in this project, we choose three aspects of data rather than only analyze performance of soccer players to analyze (compared with other researches), which is more reasonable to modeling and predicting. Besides, the method we used is related to multiple linear regression, which is easier for people to understand than using other ways. Besides, it brings a reference to soccer clubs, they should attach more importance to age, average playing time, average goals and his club rank of a player when they want to employ the player.

References

[1] E. Amir, G. Livne Accounting, valuation and duration of football player contracts Journal of Business Finance & Accounting, 32 (3–4) (2005)

ISSN 2616-7433 Vol. 2, Issue 15: 132-143, DOI: 10.25236/FSST.2020.021516

[2] T. Pawlowski, C. Breuer, A. Hovemann. Top clubs' performance and the competitive situation in European domestic football competitions Journal of Sports Economics, 11 (2), 2010.
[3] Borchers JR, Clem KL, Habash DL, Nagaraja HN, Stokley LM, BestTM. Predicting Football Players' Dual-Energy X-Ray Absorptiometry Body Composition Using Standard Anthropometric Measures, 47(3), 2012
[4] S. Herm, H.-M. Callsen-Bracker, H. Kreis. When the crowd evaluates soccer players' market values: accuracy and evaluation attributes of an online community Sport Management Review, 17 (4), 2014.
[5] Yunus Koloğlu, Hasan Birinci, Sevde Ilgaz Kanalmaz, Burhan Özyılmaz. A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position, 8(2), 2018
[6] S. Herm, H.-M. Callsen-Bracker, H. Kreis. When the crowd evaluates soccer players' market values: accuracy and evaluation attributes of an online community Sport Management Review, 17 (4), 2014.
[7] Eerika Marmo, Influence of Club Ownership Structure on Football Player Transfer Fees.11(2), 2007.
[8] V. Di Salvo, R. Baron, H. Tschan, F. J. Calderon Montero, et al. Performance Characteristics According to Playing Position in Elite Soccer; International Journal of Sports Medicine 28(3):222-7，2007.