Dean&Francis
ISSN 2959-6157

Analyzing Relationship Between Distance Covered in a Soccer Game and Defensive Performance by Regression Models

Zixuan Chen

Gaoxin No.1 High School, Xi'an, China

*Corresponding author: g20061114@gmail.com

Abstract:

Regression models are widely used in sports research, and there are mainly for two purposes: prediction and effect estimation. The article focuses on providing better strategies for soccer teams' managers based on analyzing data, which is related to the team's defense particularly. The methodology designed to find distance covered by a team as an indicator of its defensive performance. By using multiple regression models, the best fits for specific teams are created. Then based on the curve derived, the relationship between the distance covered and the players' defensive performance can be found. It is concluded that for some teams the two variables do not relate to each other. By contrast, for the teams do, though the specific relationship for each team differs slightly (they fit in different models), the defensive performance increases in intensity as distance covered increases. This detailed result may assist managers in changing tactics, injury prevention, player selection and training management.

Keywords: Regression model; Distance covered; Defensive performance.

1. Introduction

Soccer is the world's game. According to statistics, from the early 1960s to the 2018 edition of FIFA World Cup in Russia, attendance to the soccer matches has grown from 8000 to 3 million people [1]. The reason why this sport is gaining so much popularity probably is due to the highly competitive style of the game. A player averagely covers a 10km distance per game, of which 8% to 18% is at his highest speed [2]. Plenty factors can influence a soccer game, such as technical factors, mental factors, tactical factors [3]. When it comes to the tactical factors, it comes to two simple parts: how to do with the ball(attack), and how to do without the ball (defense).

Since defense is so important, it would be crucial for managers to figure out what factors could lead to defensive success. To maintain a steady defense, players must have good physical conditioning. While a player's physical conditioning stats might be hard to measure, distance covered, which greatly relates to one's physical condition, can be an appropriate substitute. Recent studies suggest that the distance covered by players, especially when pressing in high-intensity situations, correlates with a team's defensive

ISSN 2959-6157

success. An analysis of top European football clubs points out the significance of maintaining high-intensity defense by showing its link with a team's overall success. But this tendency is not true for all soccer clubs. In an article, Souza and Hoppe mention the existence of inconsistency of this link across different leagues [4].

The author first discusses the methodology used in this article--mainly regression models. It includes the procedure of the study, introduction of the regression models and relating formulas used in the study and some clarifications. Then, the author does an analysis on the result. Finally, the author mentions some useful applications of the result.

2. Method

2.1 Main Ideas and Procedures

The idea is to collect the data of distance covered for a specific team and the defensive performance, which is presented in the form of index. It then finds the relationship between the two variables. One is about distance covered, indicating total distance a team covers during matches over a season. The other is defensive performance, and this variable includes metrics like tackles, clearances and so on. These two variables can be used in different types of regression models.

The detailed procedures are the following. Firstly, using the data (in forms of coordinates) to work out the function for different regression models. Then, sketch the curve and consider the error between the predicted value and the real value to calculate R-square. Finally, compare the R-square value between different models and choose the one with the highest R-square value.

2.2 Regression Model and R-sqaure

A regression model is essentially a mathematical function that describes the relationship (quantitative) between an independent variable and a dependent variable. Regression models are used widely when the dependent and independent variables have a linear or non-linear connection, and the value of the target variable is continuous in the database. Thus, regression models can help with establishing a relationship between the two variables and also forecasting [5].

Linear Function Regression Model. Linear regression model is among the most extensively used models, which assumes a linear relationship between a dependent variable (Y) and an independent variable (X). The function of the model is defined as $Y = a + bX + e$, where ' a ' denotes the intercept, ' b ' denotes the slope of the line, and ' e ' is the error term. Sometimes linear regression model can be simple, with only one dependent and one independent variable. However, it also can be complex with numerous dependent and independent variables, or with one dependent variable and more than one independent variable.

Logistic Function Regression Model. Logistic regression model may come in handy if the dependent variable is discrete. In other words, the model can be used to compute the probability of two opposite occurrence such as true/false, 0/1, and so forth. Thus, the model has an implied limitation for the choose of the target variable: it can only be one of the two values.

Polynomial Function Regression Model. Polynomial regression is widely used when there is a non-linear relationship between the two variables. This technique is a variant of the multiple linear regression model, but the graph of the function is a curve rather than a straight line.

Exponential Function Regression Model. The exponential model is used when there is a constant growth/decline rate for the variable, such as population growth, radioactive decay and deposit in the bank. Here, it is assumed that the interest rate does not change. The model fits data to the form $y = ae^{bx}$, where a and b are constants, x is the independent variable, y is the dependent variable, and e is the base of the natural logarithm.

Power Function Regression Model. The power function regression model fits data to a relationship of the form $y = ax^{b}$, where a and b are constants, x is the independent variable and y is the dependent variable. In physics for power laws, the model is often used because in this scenario the variables involve proportional scaling. The model is suitable for a wide range of data, for it is capable of fitting in both concave data and convex data. Using the past seasons' data to create the best fit, the regression model can be used to predict the next season's defensive performance of a team corresponding to that model.

R-square can measure the accuracy of a certain model created, indicating the percentage of the variance in the dependent variable that the independent variables explain collectively. A model is more accurate when the value is closer to 1, and is less when closer to 0.

2.3 Formula and Clarifications

In the linear regression model, the R^2 is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y} - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{1}$$

where $\hat{y}$ is the predicted value generated by the model, y_i

is the value in the states, and $\bar{y}$ is the average of the stats value. For linear regression model $y = bx + a$, it is found that

$$b = \frac{n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^{2} - \sum_{i=1}^{n} x_i} \qquad (2)$$

and $a = \bar{y} - b\bar{x}$. Here, x_i and y_i are the values in the states.

Noticing other variables like tactics, play style, and league exist, the author fits the stats for different team with different models, based on which model has the highest R-square value. For example, when finding the best fit for Manchester United, after comparing different model's R-square value, the quadratic function is found out to be the best fit. But for Bayren Munich and Juventus, the best fits are linear and power, respectively. For Liverpool, the relationship between the two variables is not very significant so there is no best fit. Admittedly, the distance covered is just one of many variables that affects the defensive performance of a team, and in reality, the situation could get much more complex, the reason why distance covered is chosen is because of the following.

The first is that its Direct Influence on Defensive Actions. Distance covered reflects the physical movement for many defensive actions like intercepting, tackling and following the attackers. The second is related to a team's Fitness and Stamina. Distance covered is an indicator of a team's stamina, and defensive success highly relies on a team's ability to maintain formation organized throughout the game. Apparently, stamina is very crucial in this. The third is about Measurable. Distance covered is a more objective metric compared to other variables like fans' support and communication between players. This is exactly what makes it a more detectable and thus reliable metric to analyze.

3. Results and Applications

3.1 Four Prestigious Soccer Team

Four prestigious soccer team are analyzed in this research and before looking and the result, here are some brief introductions about the four teams [6].

Liverpool: Founded in 1892, Liverpool FC is famous for their extremely passionate fans and their efficient counter attack play style. They won the Premier League champion 19 times and UEFA champion 6 times. Anfield is where Liverpool hosts their game.

Manchester United: Established in 1878, Manchester United has won 20 English league titles and 3 UEFA champions. Known as "the red devil", Manchester United is praised for their perseverance: its players would fight for win until the very last minute. Old Trafford, its home pitch, is globally recognized.

Juventus FC: Founded in 1897, Juventus is one of the most competitive clubs in Italy, winning over 35 Serie A titles. Recognized for their steady defensive performance, Juventus FC is home to many world-class defenders like Gattuso Siria and Gianluca Pessotto.

Bayern Munich: Bayern Munich, established in 1900, is Germany's top notch soccer club, winning more than 30 Bundesliga titles and six UEFA Champions League trophies. Allianz Arena is their home pitch.

3.2 Results and Analysis

As can be seen from Fig. 1(a), generated Based on Manchester United's past 10 seasons' data in total distance covered and the defensive performance by the whole team (source: UEFA.com), the polynomial function turns out to be the best fit ($R^2 = 0.91$) compared to other functions (linear function 0.84; exponential function 0.82; logistic function 0.85; power function 0.84). The function and the graph tell that the defensive performance will first go up and then go down when the distance covered rises, and the distance that brings out the best defensive performance is 1368.5km. Such result may suggest that the Manchester United players are either wasting their energy (if distance falls above 1368.5), which could lead to fatigue and lower the team's defensive performance, or do not give their best during the game (if below 1368.5), which could also obviously lower the defensive effect.

For Liverpool (see Fig. 1(b)), Among the functions, the R-square value are all very low (linear 0.14, logistic 0.14, polynomial 0.16, exponential 0.14, power 0.14). So, it may suggest that distance covered may not be the main factor of the team's defensive performance.

This result is reasonable because in the past season, Liverpool was coached under Jurgen Klopp, a manager who prefers high-pressing play style. In that way, Liverpool's attackers account for more running distance than other teams' attackers, leading to a result that its defensive performance may be mainly affected by some other variables.

For Bayern Munich (Figure Fig. 1(c)), the best fit is linear (0.671), compared to exponential (0.67), power (0.63), logistic (0.61), and polynomial function (0.67). This suggest that more distance covered will lead to better defensive performance. However, when deploying tactics, team managers should also consider other variables as opposed to letting the team members to play in an aggressive style recklessly, since R-square is not 1. This also means there are factors influencing the defensive performance other than distance covered.

Juventus FC (Fig. 1(d)) is an interesting case. When analyzing the past 10 seasons' data as a whole, one tends to conclude a tenuous relationship between the two variables. The line generated based on the blue dot and the 9 orange dots, with R-square only 0.14. But after excluding the data in season 2022/23 (corresponding to the point (105.4,68) in Fig. 1(a)), the model shows a relatively strong relationship between two variables. They are generated based on the 9 orange dots, with R-square being 0.56.

This could also suggest that distance covered is a strong indicator of the defensive performance by Juventus FC, and what happened in 2022/23 season could just be an exception. And after noticing this, the author delves into the club's news, discovering that while Massimiliano Allegri has been the club's manager since 2014, he is replaced by Thiago Motta in 2023.Undoubtedly, a new coach will surely result in changes in the team's original play style, which could account for the excepted point.

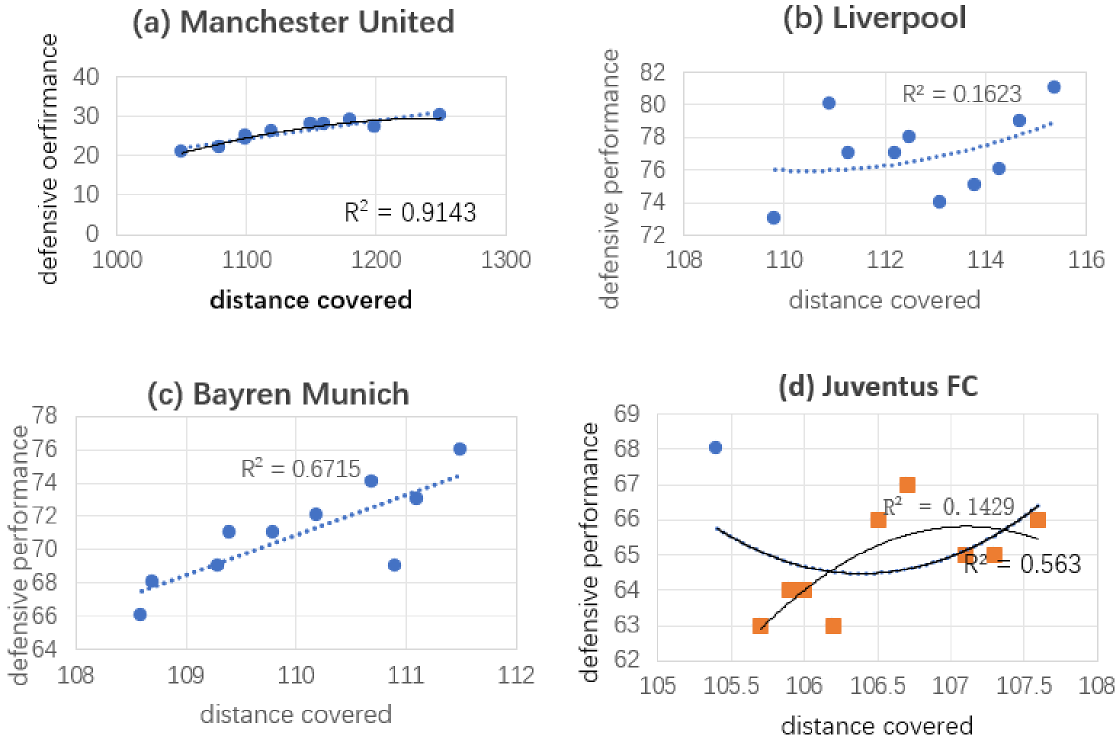

Fig. 1 The defensive performance vs distance coverd for four different teams.

3.3 Useful Applications

3.2.1 Injury Prevention and Load Management

The model can be used to prevent injuries by giving team manager a signal when to change the play-style in the middle of a game. The manager can deploy tactics, which can interfere with the index of defensive performance, during the half-time break based on the distance covered by players in the first half. In doing so, it might affect the distance covered in the second half and prevent injuries. For example, if the model suggests that higher defensive actions lead to excessive distance covered that could lead to fatigue and possible injuries, the manager could make decision to change to a less aggressive pressing play-style if there is a big game coming up and he wants the team members to conserve energy for it.

3.2.2 Strategic Match Planning

The model can also be used for managers to analyze their opponents. With the understanding of the relationship between the distance covered and the defensive performance, the managers can analyze their opponent's defensive intensity, which has to do with a team's play-style, by analyzing the distance covered by that team. And after knowing the habits of their opponents, the manager can tailor the tactics for the match against that particular team. For example, when playing against a high-pressing team like Liverpool, the manager might choose a tactic like a distance range that while ensuring the defensive performance. Thus, they can try to conserve as much energy as possible for the counter attack [7]. For teams like

Manchester United, the model shows that their defensive performance reaches the maximum when the team covers certain amount of distance during a season and drop when the distance goes either up or down. This may be telling the managers that the excessive running distance is superfluous: it overburdens the player so that their defensive performance reduces because of fatigue. To cope with that, the managers might buy in new players with better stamina to balance the situation, or focus more on physical training, particularly cardio, to improve on the team's stamina. He could also tailor new tactics based on the "best distance" that shown in the graph, letting the team to play in a more/less aggressive style so that the distance covered is increased/reduced, which leads to the optimal balance by reaching the distance covered that brings out the best defensive performance.

3.2.3 Other Examples

Teams aiming to improve their defensive performance could consider adding more cardio training to their training programs, since more distance covered leads to better defensive performance. The relationship between the distance covered and defensive performance highlights the importance of stamina as a trait when coaches are looking for a good defender. When buying new players during the transfer window, coaches can regard the data of the player's covered distance as the reflection of his defensive ability.

4. Conclusion

In the article, the author introduces the regression models and how they can be used to analyze the relationship between the two variables: distance covered and defensive performance. The result shows that for some teams the two variables do not relate to each other. However, for the teams do, though the specific relationship for each team differs slightly that fit in different models, the defensive performance increases in intensity as distance covered increases. Such knowledge brings more insight on evidence-based adjustment to soccer teams. If a team knows their relationship of the two variables, they can tailor their tactics, player selection and training focus based on the relationship. However, the accuracy of the model still needs improvement. By now The Value-Defensive Event Prediction model is the prevailing metric to assess defensive performance by a team. Nevertheless, it is less effective in evaluating rare events such as goals. This suggests that future studies should focus on more frequent events like ball recoveries or successful ball interceptions to better classify defensive performances. Except the metric for defensive performance, the precision of measuring distance covered also need enhancement. Inconsistent data from GPS systems and the challenge of accurately recording player movements at high intensity, especially during deceleration and acceleration, can limit the accuracy of distance covered recorded in match play.

References

[1] Yamashita, Gabrielli H., et al. Customized prediction of attendance to soccer matches based on symbolic regression and genetic programming. Expert Systems with Applications, 2022, 187: 115912.

[2] Ekblom, Björn. Applied Physiology of Soccer. Sports Medicine, 1986, 3(1): 50–60.

[3] Stølen, Tomas, et al. Physiology of Soccer. Sports Medicine, 2005, 35(6): 501–36.

[4] Modric, Toni, et al. Analysis of Running Performance in the Offensive and Defensive Phases of the Game: Is It Associated with the Team Achievement in the UEFA Champions League? Applied Sciences,2021, 11(18): 8765.

[5] Toda, Kosuke, et al. Evaluation of Soccer Team Defense Based on Prediction Models of Ball Recovery and Being Attacked: A Pilot Study. PLOS ONE, 2022, 17(1): e0263051.

[6] Harper, Damian J., et al. High-Intensity Acceleration and Deceleration Demands in Elite Team Sports Competitive Match Play: A Systematic Review and Meta-Analysis of Observational Studies. Sports Medicine (Auckland, N.Z.), 2019, 49(12): 1923-47.

[7] Wiedermann, Wolfgang, et al. Heteroscedasticity as a Basis of Direction Dependence in Reversible Linear Regression Models. Multivariate Behavioral Research, 2017, 52(2): 222–41.