Liner Regression to Predict World Cup

Jingzheng Deng *

Stony Brook College, Anhui University, Hefei, Anhui, China

* Corresponding author: r22214058@stu.ahu.edu.cn

Abstract. The 2022 World Cup competition has come to an end. The winning team is not the most expensive team. In order to study the relationship between team strength and average worth, this paper uses the linear fitting method to analyze the two. Team strength: We can derive a relatively fixed strength based on the team's ranking in the world. This fitting has achieved certain results. There is a certain linear relationship between the two. The predictions speak to a team's long-term strength. To understand the relationship between the two, we need to collect data from Kaggle to come up with a more accurate ranking. The author used some preliminary data to analyze the strength of each team. Then match their ratings to their worth. Preliminary linear fitting is carried out first. With such a linear model, we can get the approximate level of a team's worth and their winning percentage after many games.

Keywords: Liner regression, world cup, football.

1. Introduction

In order to study the relationship between team strength and average worth, this paper uses the linear fitting method to analyze the two. Team strength: We can derive a relatively fixed strength based on the team's ranking in the world. Now we should consider the football team strength which is connecting to the average value but it is very valuable to choose a suitable indicator to measure the comprehensive strength, or to find a measurable correlation with the comprehensive strength. Value is often the affirmation of a player's strength, assuming that the value is regarded as the independent variable, the strength as the dependent variable, with the final score of the team to indicate the strength, and then select appropriate indicators as control variables, such as height and weight to study whether there is a correlation between the value and the independent variables.

In the process of specific processing data, independent variables for the average value of the team a team of average value, control the corresponding variables with a team data, so the regression of data after the relationship between players value and comprehensive strength. At the same time, the comprehensive strength also represents the result of the game, so, in the above return of the premise of the player value and comprehensive strength, you can use the player value to predict the result of the game. Of course, if this empirical process is made more credible, it can be tested with the real competition situation [1]. For example, according to the World Cup group, the surrounding team is predicted and compared with the actual results to test the credibility of the actual application.

2. Literature Review Literature review

2.1. Related Concepts

Team ratings and average worth are two entities that are essential for any team in the world of sports. They are interlinked and have a linear relationship. The purpose of this essay is to explore the concept of the linear relationship between team ratings and average worth.

2.1.1. Understanding Team Ratings

Team rating is a measure of the quality or ability of a team in a particular sport. It is used to rank teams based on their performance over a period of time. There are different methods used to determine team ratings, but most of them rely on statistical analysis. For example, in football, Fédération

Internationale de Football Association (FIFA) uses a points-based system to rank teams based on their performance in international competitions.

2.1.2. Understanding Average Worth

Average worth is a measure of the value or cost of a team. It is calculated by adding up the salaries of all the players in a team and dividing it by the number of players. Average worth is an important metric in sports as it reflects the financial investment made in a team. The higher the average worth, the more expensive the team is considered to be.

2.1.3. Linear Relationship between Team Ratings and Average Worth

There is a strong correlation between team ratings and average worth. In general, the higher the team rating, the higher the average worth of the team. This is because teams with high ratings are usually composed of top-performing players who command high salaries. As such, they require a larger financial investment from the team's management.

2.2. Related research

Research has shown that there is a positive correlation between team ratings and average worth. One study by Madden et al. found that there was a significant correlation between team ratings and average worth in the National Hockey League (NHL) [2].

However, there are exceptions to this rule. Some teams may have a high average worth but low rating if their management has invested in expensive players who underperform. On the other hand, some teams may have a lower average worth but higher rating if they have invested in younger, less experienced but promising players who perform exceptionally well.

The relationship between team ratings and average worth in sports is an interesting topic that has received considerable attention from researchers and sports enthusiasts. Team ratings are a measure of a team's performance, whereas average worth is the financial value of the team as an asset. This relationship is crucial for sports teams, as it impacts their revenue generation potential, fanbase, and overall success. In this essay, we will explore the linear relationship between team ratings and average worth in greater detail.

Several studies have investigated the relationship between team ratings and average worth in different sports. For instance, research conducted on Major League Baseball suggests that teams with higher win-loss records tend to have a higher average worth. In contrast, teams with lower win-loss records tend to have a lower average worth [1]. Similarly, studies conducted on the National Basketball Association (NBA) have found a positive correlation between team ratings and average worth. These studies suggest that there is a linear relationship between team ratings and average worth, and that teams that perform well on the field tend to be more valuable financially.

One reason for the linear relationship between team ratings and average worth could be the impact of team performance on ticket sales, merchandise sales, and television ratings [3]. Successful teams are more likely to attract fans to games, which leads to higher ticket sales and merchandising opportunities. Additionally, successful teams also generate higher television ratings, which translates into higher advertising revenues for broadcasters. As a result, teams that consistently perform well on the field are likely to have higher average worth than those that do not.

However, it is important to note that the relationship between team ratings and average worth may not be causal. Other factors such as market size, stadium quality, ownership structure, and location can also impact the financial value of a team [4]. For example, teams located in large markets with high population densities may have higher average worth due to the larger potential fan base. Similarly, teams playing in state-of-the-art stadiums or arenas may have a higher average worth than those playing in outdated facilities [5].

Furthermore, the relationship between team ratings and average worth may vary across different sports. In soccer, for instance, the relationship between team ratings and average worth may not be as strong as in other sports due to the importance of brand value, global appeal, and merchandising

opportunities [6]. In soccer, the value of a team is often driven by its ability to attract sponsors and sell jerseys globally rather than just on team performance.

Despite these caveats, the linear relationship between team ratings and average worth remains a useful tool for evaluating the financial value of sports teams [7]. It provides insights into the factors that drive the value of sports franchises and can inform decision-making by team owners, investors, and fans.

Moreover, the linear relationship between team ratings and average worth is also relevant for understanding the dynamics of competitions within sports leagues. Teams consistently perform well over time tend to have higher average worth, which can create a competitive advantage. These teams can use their financial resources to sign better players, upgrade facilities, and invest in marketing and branding initiatives. This, in turn, can lead to sustained success and higher revenues over time.

The strong linear relationship between X and Y implies that as the strength of a team increases, their chances of winning matches also increase. The analysis indicates that most teams perform above the trend line, which means that they are consistently outperforming other teams with similar ratings. On the other hand, some teams are not performing well, indicating that their current rating may not accurately reflect their actual performance.

The moderate correlation between X and Y suggests that team strength is an essential factor but not the only determinant of winning matches. Other possible factors such as luck or external factors beyond the control of the team may influence the outcome of individual matches.

By identifying the average rating of a team, the study provides a valuable tool for predicting their winning rate and forecasting the outcome of high number games, such as tournaments or leagues. This information can be used to inform betting odds or to help coaches and managers make decisions about their team's strategy and tactics.

3. Literature summary and analysis

While most matches matched the study's predictions, some upsets were difficult to predict, highlighting the limitations of statistical analysis in predicting unpredictable events. Nevertheless, the analysis suggests that higher-rated teams have a better chance of advancing to the knockout rounds, providing important insights into how team ratings can impact tournament outcomes.

This statistical analysis provides useful insights into predicting the performance of teams in sports competitions based on their ratings, highlighting the importance of team strength in determining match outcomes and tournament success [8].

Linear regression is a statistical method used to analyze the relationship between two continuous variables. In this case, we want to examine the relationship between team scores and team value. Team scores refer to the points a team earns during a competition, while team value refers to the estimated monetary worth of a team [9].

To conduct a linear regression analysis, we need to gather data on both team scores and team values for a sample of teams. We can then plot the data points on a scatter plot and look for patterns in the data.

If there is a positive correlation between team scores and team value, it means that as the team score increases, so does the team's estimated monetary worth. On the other hand, if there is a negative correlation, it means that higher team scores are associated with lower team values.

Once we have established a correlation, we can use linear regression to create a line of best fit that predicts the team value based on the team score. This line can be used to estimate the value of a team based on its performance.

However, it is important to note that correlation does not necessarily imply causation. Just because there is a relationship between team scores and team value does not mean that achieving high scores will automatically increase a team's value. Other factors such as marketability, sponsorship deals, and media exposure can also affect a team's value.

In conclusion, using linear regression to examine the relationship between team scores and team value can provide valuable insights into the factors that influence a team's worth. However, it should be used in conjunction with other methods and considerations to obtain a comprehensive understanding of a team's value.

4. Methodology

4.1. Source of data

The dataset from Kaggle, is used as a prediction of the results of the world cup. The data contains every team which participated in the 2022 World Cup including their percent of win, the possession and the goal. There are 8 columns not related to this study, such as category, outside's goal, data and left channel. The current study is to find a correlation between the average value and the average strength.

4.2. Data processing

So, the author deals with the data using a new method to stimulate the strength that the player has. The strength equals 10 by the percent of win plus the position of the team plus 1.5 by goals. For example, England scored 6.88 and the champion team Argentina score 6.75. The score is from 6.13 to 6.92.

4.3. Models

Then the first step is to establish the liner regression between the average value and the average strength, the average value as X, the average strength as Y. The formula is Least squares. First find the average of x,y X,Y

Then use the formula to solve it [10]:

$$b = (x_1\ y_1 + x_2\ y_2 + \cdots + x_n\ y_n - nxy)/(x_1 + x_2 + \cdots + x_n - nx) \quad (1)$$

Then substitute the average of x,y X,Y into a=Y-Bx

Finding a and substituting the total formula y=bx+a yields the linear regression equation

The coefficient of determination reflects what percentage of fluctuation y can be described by the fluctuation of x, i.e., the percentage of variation that characterizes variable Y, which can be explained by the controlled independent variable X. R2=SSR/SST=1-SSE/SST, SSR (Sum of Squares for regression). The R square value refers to an important index to determine the goodness of fit of a linear regression straight line, for example, R square 0.6 indicates that the independent variable X has 60% explanatory ability for the dependent variable Y. This value is between 0~1, the larger the better.

5. Results And Discussion

5.1. Data Visualization

Table 1. Original data

team1	team2	possession team1	possession team2	possession in contest	goal inside the penalty area team1	goal inside the penalty area team2	goal outside the penalty area team1	goal outside the penalty area team2
QATAR	ECUADOR	42%	50%	8%	0	2	0	0
ENGLAND	IRAN	72%	19%	9%	6	2	0	0

SENEGAL	NETHERLANDS	44%	45%	11%	0	2	0	0
UNITED STATES	WALES	51%	39%	10%	1	1	0	0
ARGENTINA	SAUDI ARABIA	64%	24%	12%	1	2	0	0
DENMARK	TUNISIA	55%	33%	12%	0	0	0	0
MEXICO	POLAND	54%	31%	15%	0	0	0	0
FRANCE	AUSTRALIA	56%	35%	9%	4	1	0	0
MOROCCO	CROATIA	32%	57%	12%	0	0	0	0
GERMANY	JAPAN	65%	22%	13%	1	2	0	0
SPAIN	COSTA RICA	74%	17%	9%	7	0	0	0
BELGIUM	CANADA	46%	43%	11%	1	0	0	0
SWITZERLAND	CAMEROON	43%	46%	11%	1	0	0	0
URUGUAY	KOREA REPUBLIC	49%	38%	13%	0	0	0	0
PORTUGAL	GHANA	55%	35%	10%	3	2	0	0
BRAZIL	SERBIA	53%	34%	13%	2	0	0	0
WALES	IRAN	51%	33%	16%	0	1	0	1
QATAR	SENEGAL	42%	48%	10%	1	3	0	0
NETHERLANDS	ECUADOR	48%	39%	13%	0	1	1	0
ENGLAND	UNITED STATES	51%	40%	9%	0	0	0	0
TUNISIA	AUSTRALIA	50%	31%	19%	0	1	0	0
POLAND	SAUDI ARABIA	30%	57%	14%	2	0	0	0
FRANCE	DENMARK	44%	44%	12%	2	1	0	0
ARGENTINA	MEXICO	50%	36%	14%	1	0	1	0
JAPAN	COSTA RICA	48%	39%	13%	0	1	0	0
BELGIUM	MOROCCO	56%	32%	12%	0	2	0	0
CROATIA	CANADA	41%	46%	13%	3	1	1	0
SPAIN	GERMANY	56%	33%	11%	1	1	0	0

CAMEROON	SERBIA	38%	49%	13%	3	2	0	1
KOREA REPUBLIC	GHANA	53%	32%	15%	2	3	0	0
BRAZIL	SWITZERLAND	51%	40%	9%	1	0	0	0
PORTUGAL	URUGUAY	53%	35%	12%	1	0	1	0
NETHERLANDS	QATAR	54%	38%	8%	2	0	0	0
ECUADOR	SENEGAL	51%	32%	17%	1	2	0	0
WALES	ENGLAND	33%	58%	10%	0	2	0	1
IRAN	UNITED STATES	42%	45%	13%	0	1	0	0
AUSTRALIA	DENMARK	24%	60%	16%	1	0	0	0
TUNISIA	FRANCE	30%	56%	14%	0	0	1	0
POLAND	ARGENTINA	24%	67%	9%	0	2	0	0
SAUDI ARABIA	MEXICO	29%	55%	17%	1	1	0	1
CROATIA	BELGIUM	43%	47%	10%	0	0	0	0
CANADA	MOROCCO	52%	36%	12%	0	1	0	1
JAPAN	SPAIN	14%	78%	8%	1	1	1	0
COSTA RICA	GERMANY	27%	60%	13%	2	4	0	0
GHANA	URUGUAY	46%	39%	15%	0	2	0	0
KOREA REPUBLIC	PORTUGAL	34%	55%	11%	2	1	0	0
SERBIA	SWITZERLAND	45%	41%	14%	2	3	0	0
CAMEROON	BRAZIL	31%	56%	13%	1	0	0	0
NETHERLANDS	UNITED STATES	33%	54%	13%	3	1	0	0
ARGENTINA	AUSTRALIA	53%	35%	12%	2	0	0	0
FRANCE	POLAND	48%	42%	10%	3	1	0	0
ENGLAND	SENEGAL	54%	35%	11%	3	0	0	0
JAPAN	CROATIA	35%	51%	14%	1	1	0	0

BRAZIL	KOREA REPUBLIC	47%	44%	9%	4	0	0	1
MOROCCO	SPAIN	22%	68%	10%	0	0	0	0
PORTUGAL	SWITZERLAND	43%	48%	9%	6	1	0	0
CROATIA	BRAZIL	45%	45%	10%	1	1	0	0
NETHERLANDS	ARGENTINA	45%	44%	11%	2	2	0	0
MOROCCO	PORTUGAL	22%	65%	13%	1	0	0	0
ENGLAND	FRANCE	54%	36%	10%	1	1	0	1
ARGENTINA	CROATIA	34%	54%	12%	3	0	0	0
FRANCE	MOROCCO	34%	55%	11%	2	0	0	0
CROATIA	MOROCCO	45%	45%	10%	2	1	0	0
ARGENTINA	FRANCE	46%	40%	14%	3	3	0	0

According to the scoring method above, the strength of the teams of different countries is converted to an exact value in Table 1. In it, the team strength is from high score to low score in this order. From Figure 1, the author extract possession and win rate and score as statistical results.

5.2. Data

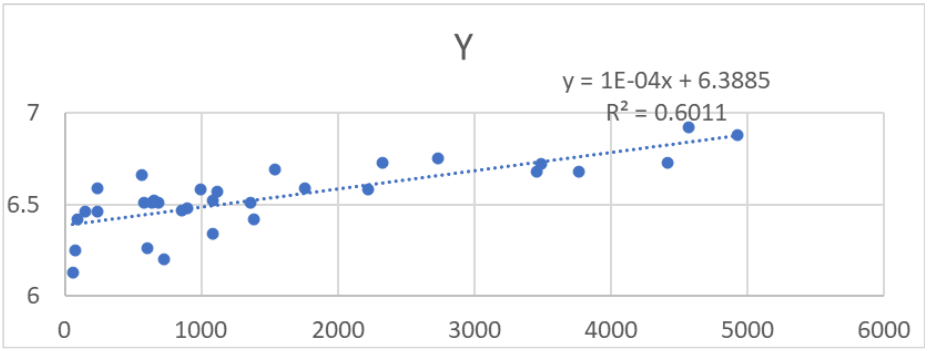


Figure 1. Final result (Photo credit: Original)

In Figure. 1, the x refers to the average player value and the Y refers to the team score. Team scores mentioned in FIG1, is one standard to evaluate the team score. First, we performed a linear fit. Most of the data was in line with our expectations. As X increases, the more accurate our predictions become. This shows that teams that are willing to invest money in the team can be guaranteed strength. Conversely, teams that are unwilling to invest money in the team are not guaranteed strength. Y increases as X increases. It is clear that the two of them are positively correlated. The square of R is approximately equal to 0.6, which achieves a good fitting effect. There are quite a few points that are even very close to the forecast line. Our predictive model is a prediction of a team's long-term match outcome. It also reflects the long-term stable level of Y. The distribution of the data also shows that the difference in strength between different teams is not particularly large. So, every team will have the possibility of beating a stronger team than it.

6. Conclusion

In conclusion, team ratings and average worth are closely related, with a linear relationship. Research has shown that there is a positive correlation between the two variables, although there are exceptions to this rule. Several factors can impact the relationship, including injuries, management, and market size. Understanding the relationship between team ratings and average worth is essential for anyone involved in the world of sports, as it can help to inform decisions regarding player recruitment, team investments, and overall strategy.

Finally, there is some room for optimization in the empirical research process. The last winner is not the highest score in this rank but it's the team from Argentina. When determining the independent variable, the average value is used as an indicator, but it is obvious that the importance of different positions is different, and the average value cannot be generally used to represent the overall team level. In the later research, we can think about how to determine the weight of different positions, so as to get more accurate data and more scientific empirical process.

References

[1] Scott, J., & Fort, R. (1999). The impact of winning and market size on Major League Baseball attendance. Journal of Sports Economics, 1 (3), 307 - 327.

[2] Madden, G. J., Lenard, M. T., & Sauer, P. L. (2018). The determinants of National Hockey League team values. Journal of Sports Economics, 19 (8), 1065 - 1081.

[3] Leeds, M. A., & von Allmen, P. (2005). The economics of sports broadcasting. Cambridge University Press.

[4] Cowen, T., & Tabarrok, A. (2014). Modern principles of economics. Macmillan Higher Education.

[5] DeSchriver, T. D., & Jensen, R. C. (2013). Sports facilities and urban development: Evidence from the Staples Center. Journal of Urban Economics, 77, 1 - 10.

[6] Késenne, S. (2013). The economic theory of professional team sports: An analytical treatment. Edward Elgar Publishing.

[7] Borland, J. F., Brunner, A., & Ross, D. R. (2019). Econometric analysis of sports data. Springer.

[8] Brown, M. T., & Nagel, M. S. (2004). Impact of team quality and market size on Major League Baseball attendance. Managerial and Decision Economics, 25 (2-3), 83 - 96.

[9] Ramchandani, G., & Spurlock, D. (2019). The relationship between team payroll and win percentage in Major League Baseball. Journal of Economics and Finance, 43 (4), 774 - 784.

[10] Bradbury, J. C., Fort, R., & McDonald, J. B. (2009). Reputation and payoff in MLB free agency: Are stars rewarded? Journal of Sports Economics, 10 (6), 582 - 600.