Total Team Spending and Profitability Results in a Higher League Ranking in the English Premier League*

An Analysis of Economic Predictors on League Rank in the 2023-2024 Season

Isabella MacLean

April 4, 2024

This project will investigate how to formulate a linear regression model based on three explanatory variables: match attendance, total payroll, and market value, in order to try and predict a teams finishing place in the standings. Utilizing R's lm function, this paper analyzes the relationship between these variables and teams total points accumulated during the 2023-2024 season. Findings suggest that each explanatory variable significantly influences a team's league position (total points), highlighting the association between team success and team wealth.

Table of contents

*Code and data are available at: https://github.com/Bellamaclean7/English_Premier_League_Economic_Predictors_on_League_Rank_2023-2024_Season.git.

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019).

The remainder of this paper is structured as follows. Section 2....

2 Data

3 Model

3.1 Linear Regression

A linear regression model is a statistical model for linear relationships between variables given by

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + \varepsilon_i \quad \text{for } i = 1, \dots, n. \tag{1}$$

As shown in equation (1) a linear regression model has the following components:

- Y_i: The dependent variable for the i^{th} observation. This is the response variable that is trying to be predicted or explained, dependent on the corresponding explanatory variables $X_{i,1}, X_{i,2}, \dots, X_{i,p}$.
- β_0: The intercept of the regression line. It represents the expected value of Y_i when all the independent variables (X) are equal to 0.

- $\beta_1, \beta_2, \ldots, \beta_p$: The coefficients of the model. Each β_j (for $j = 1, \ldots, p$) represents the expected change in Y_i for a one-unit change in the j^{th} independent variable, $X_{i,j}$, holding all other variables constant.
- $X_{i,1}, X_{i,2}, \ldots, X_{i,p}$: The independent variables (also called predictors or explanatory variables) for the i^{th} observation. These variables are used to predict the value of the dependent variable.
- ε_i: The error term for the i^{th} observation. It represents the difference between the observed value of the dependent variable and the value predicted by the model. It's assumed to be randomly distributed with a mean of 0.
- $i = 1, \ldots, n$: This indicates that the equation applies to each observation in the dataset, from the first ($i = 1$) to the n^{th} (the last observation), where n is the total number of observations.

3.2 Model set-up

$$\begin{aligned} \text{pts}_i = {} & \beta_0 + \beta_1 \times \text{average\_home\_matchday\_attendance}_i \\ & + \beta_2 \times \text{total\_wage\_bill}_i + \beta_3 \times \text{market\_value}_i + \varepsilon_i \end{aligned} \tag{2}$$

Equation (2) can be explained as follows:

- pts_i: The dependent variable for the i^{th} observation. This is the outcome variable that the model is trying to predict or explain, which in this context could represent points (or any other metric of success) associated with each observation.
- $\beta_1, \beta_2, \beta_3$: The coefficients of the model. Each of these coefficients represents the expected change in pts_i for a one-unit increase in their respective independent variable, assuming all other variables are held constant.
  - β_1 is associated with $\text{average\_home\_matchday\_attendance}_i$, indicating how changes in home matchday attendance are expected to affect pts_i.
  - β_2 corresponds to total_wage_bill_i, reflecting the impact of the total wage bill on pts_i.
  - β_3 is linked with market_value_i, showing how the market value is predicted to influence pts_i.

3.3 Model justification

The linear regression model given by Equation (2), derived from the foundational principles of linear regression outlined in Equation (1), provides a transparent mechanism to quantify the impact of various factors on the total points accumulated by EPL teams (a direct correlation to team rank in league standings). The model will aid in gaining insights into how attendance, total payroll, and market value could affect performance metrics. The model is operationalized using the `lm` function in `R`.

4 Results

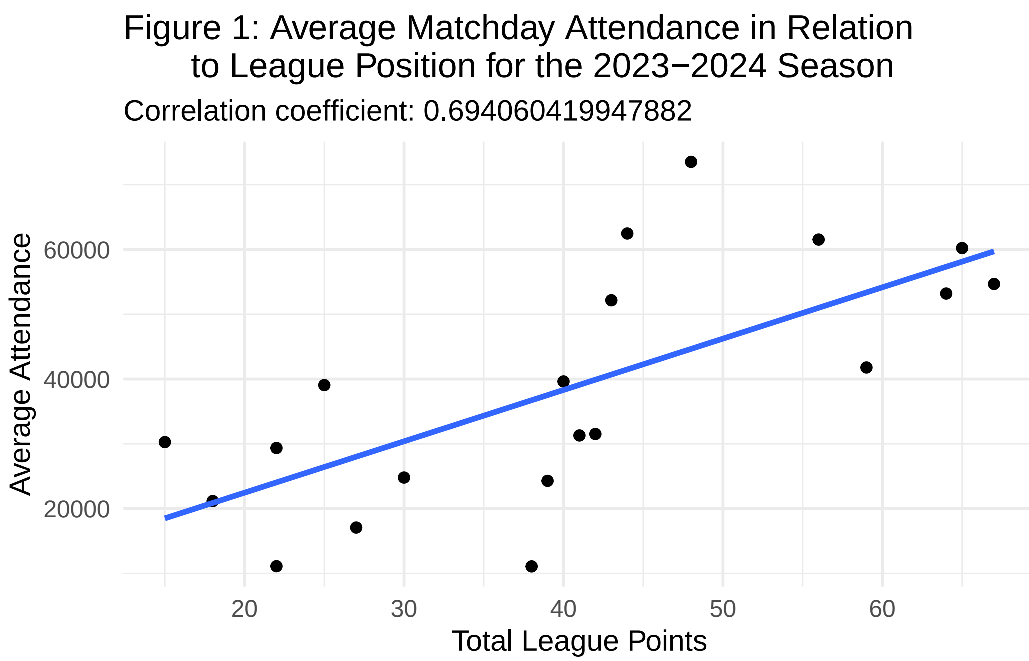

Figure 1: Figure 1 presents a scatter plot illustrating the relationship between total league points and average matchday attendance during the 2023-2024 season. Each point represents a team, plotted with their total points on the x-axis and their average home attendance on the y-axis. The trend line, added using linear regression analysis, suggests a correlation between team performance and matchday turnout. The correlation coefficient of 0.694060419947882 indicates a positive association with moderate correlation, where teams with higher league points tend to have greater average attendance.

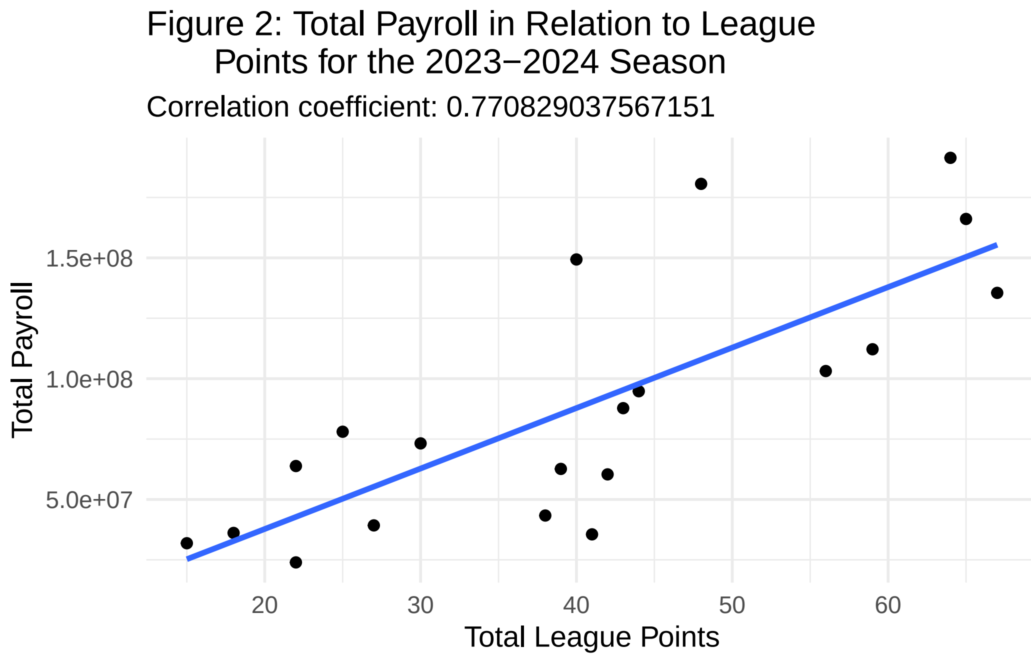

Figure 2: Figure 2 depicts the association between total league points and the total payroll for teams in the 2023-2024 season. Each point on the plot corresponds to a team, positioned according to their accumulated points for the season along the x-axis and their respective payroll on the y-axis. The linear regression trend line suggests a correlation between a team's financial expenditure on payroll and their success in the league. A correlation coefficient of 0.770829037567151 reflects a positive relationship, implying that teams with higher payrolls may be more likely to accumulate greater league points.

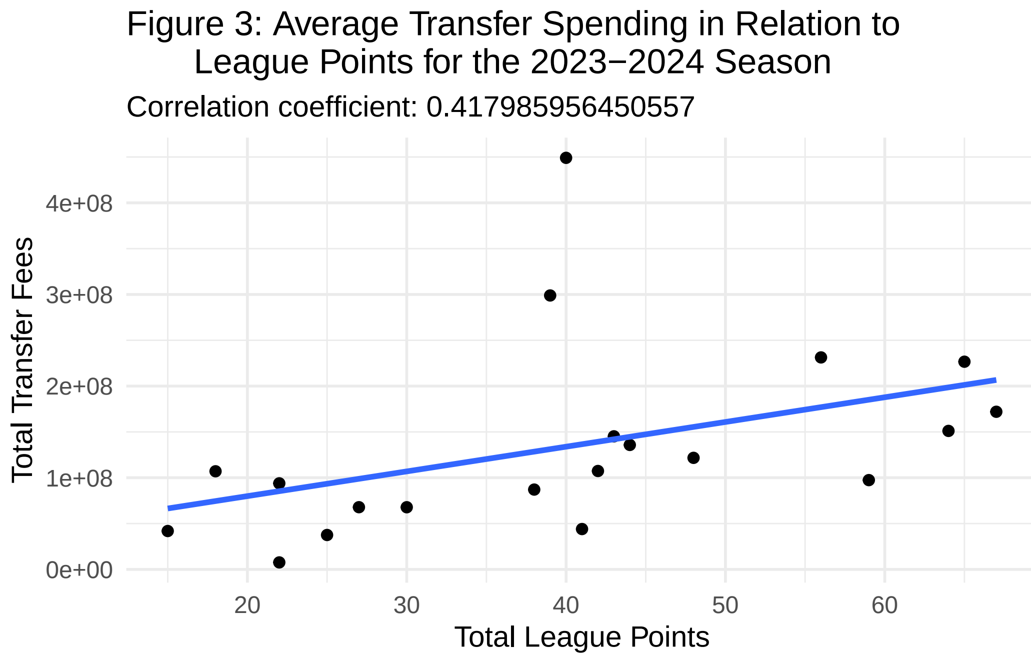

Figure 3: Figure 3 illustrates the relationship between teams total transfer spending and their corresponding league points for the 2023-2024 season. The scatter plot shows each team as a distinct point, positioned by their total league points on the x-axis and their transfer spending on the y-axis. The trend line, calculated with linear regression, indicates a moderate positive correlation, with a correlation coefficient of approximately 0.417985956450557. This suggests that higher league points tend to coincide with increased transfer spending, although the relationship is not strongly linear.

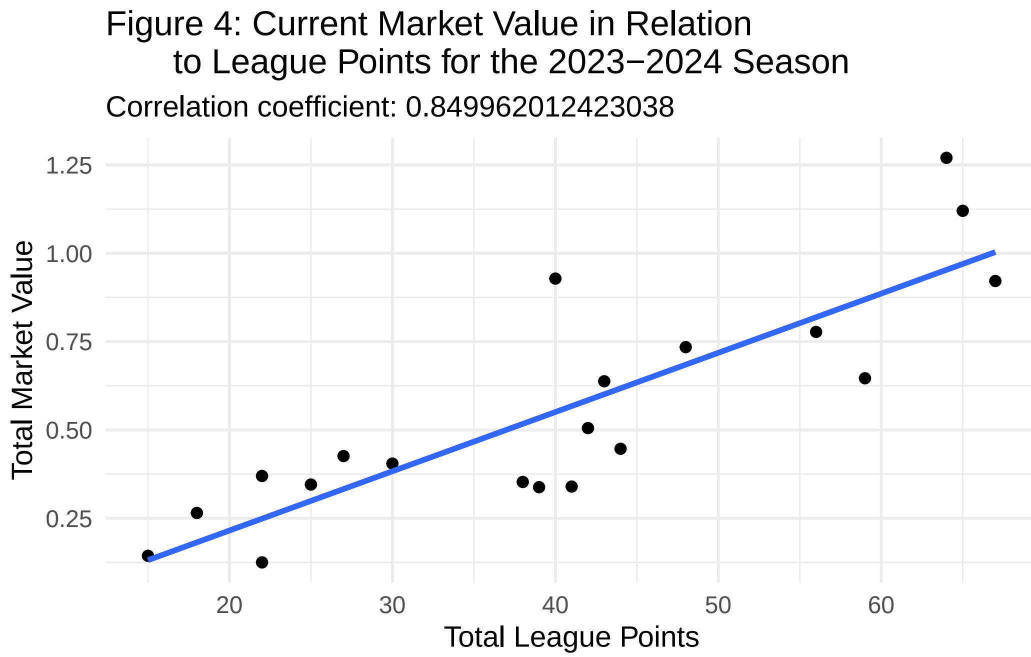

Figure 4: Figure 4 demonstrates the correlation between the current market value of teams and their accumulated league points for the 2023-2024 season. The scatter plot places each team by the total points they have secured on the x-axis against their market value on the y-axis. A pronounced upward trend represented by the linear regression line, with a correlation coefficient of 0.849962012423038, indicates a strong positive relationship. This suggests that teams with higher market values tend to amass more league points, underlining a potential link between financial strength and on-field success.

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

B.2 Diagnostics

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.