

Assignment #3: Pythagorean Theorem

Introduction and Methodology

This report looks to analyze whether score-differential is a strong predictor of future team success in soccer through usage of the Pythagorean formula, where, adjusted for ties,

$$\text{Percent of Points Gained}(\text{Point}\%) = \frac{1}{1 + (\text{Goals Scored}/\text{Goals Conceded})^{-\alpha}}$$

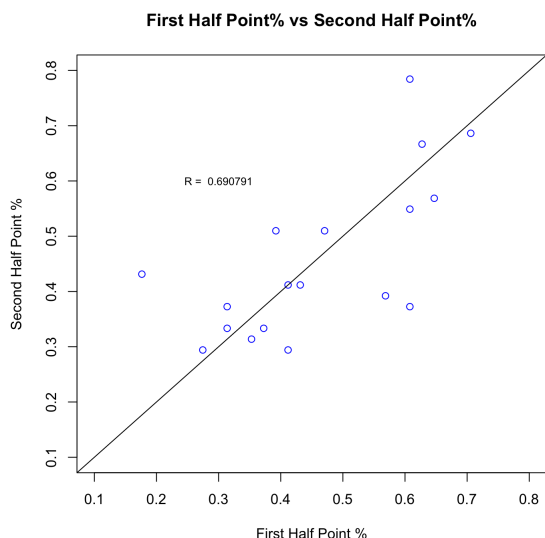
Specifically, I looked into the 2022-2023 Bundesliga Season, getting game data from fbref.com by using the [worldfootballR](https://www.tidyverse.com/packages/foreign/) R package, amalgamating the game data to get each team's goal scored and conceded as well as the percentage of points gained out of total points.

Using the data, I found an appropriate α to use in the Pythagorean formula by finding an α which minimized the sum of the residuals squared between the Pythagorean Point% and the actual Point%. After multiple iterations, I found the $\alpha = 1.3513$, with an error of **0.0546**.

Analysis

To compare the effectiveness of the Pythagorean Theorem, I then split the Bundesliga dataset into two equal subsets, with games before and after January 25, 2023, comparing the correlations between first and second-half observed Point%, first-half Pythagorean Point% and second-half observed %, and first and second-half residuals between Pythagorean Point% and Observed Point%.

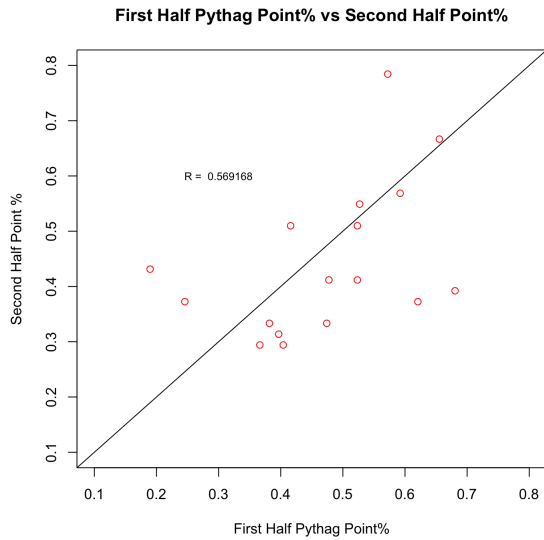
First-Half Actual Point% and Second-Half Actual Point%



Based on the relatively high correlation(**0.6907**), I can infer that a team's first-half actual performance is a strong predictor of second-half actual performance, meaning that a team's previous Point% can be a strong estimation of a team's future success. Teams who consistently win, either by one goal or many goals, are expected to continue to win throughout the season, and vice-versa.

This correlation also means that teams who have over-performed or under-performed their preseason expectations will usually continue their performance. A Cinderella story who has started the season very well will not suddenly drop off to their pre-season expectation, and an under-performing highly rated team will not suddenly start winning.

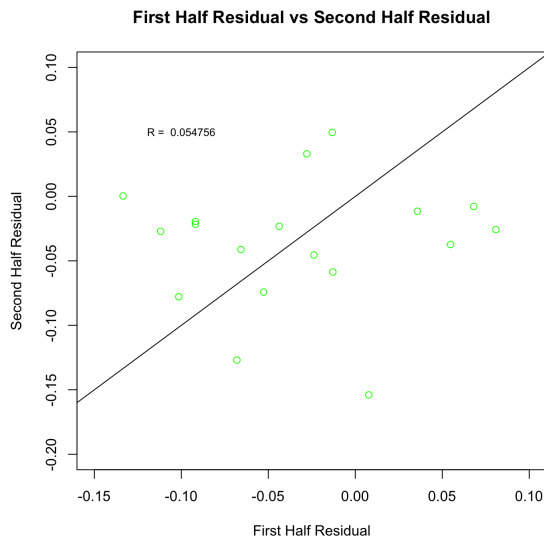
First-Half Pythagorean Point% and Second-Half Actual Point%



While still having a relatively strong correlation(**0.5668**), the relationship was notably weaker than the relationship between a team's first and second-half observed Point%, meaning that score differential is not as strong as a predictor in a team's true ability than a team's results.

This difference may be a result of the fact that a team's true ability may not be best measured through goal differential because of the difficulty to score goals. A team could dominate multiple games but only win by one goal, leading to goal differential being an unoptimal measure of true ability.

First-Half Residual% and Second-Half Residual%



Furthermore, there was very little evidence to support the fact that a team's over or underperformance from their Pythagorean expectation in the first-half has little correlation with their over/under-performance in the second-half because of the correlation(**0.0852**) between first and second half residuals.

A team who under or over-performs from their Pythagorean expectation in the first-half of the season is not likely to continue that trend in the second-half of the season.

However, it should be noted that because of the small sample size of just one season, the data is very prone to outliers, meaning that if one team drastically over-performs their Pythagorean expectation and then regresses, the correlation would be effected heavily.

Conclusion

Through the Pythagorean formula, this report looks to determine whether score-differential or the percentage of points gained was a better predictor of future team success. Although both score-differential and point percentage were strong predictors, point percentage was notably a stronger predictor. Furthermore, I found that teams who under or over-perform their Pythagorean expectation are not expected to continue this deviation in the future.