Caution: These lecture notes are under construction. You may find parts that are incomplete.

2 Pythagorean Formula

2.1 Introduction

The Pythagorean formula, first proposed by Bill James in ????, estimates a baseball team's winning percentage (W%) as a function of the number of runs scored (RS) and the number of runs allowed (RA):

$$\hat{\text{W\%}}_{\text{Pythag}} = \frac{\text{RS}^2}{\text{RS}^2 + \text{RA}^2}.$$

The name "Pythagorean formula" is a reference to the Pythagorean theorem from geometry, which asserts that $a^2 + b^2 = c^2$, where c is the length of the hypoteneuse of a right triangle, and the lengths of the other two sides are a and b. The Pythagorean formula of sport analytics is not a theorem, but it does bear some resemblance to the Pythagorean theorem of geometry.

The key observation justifying the use of the Pythagorean formula is that a team's run differential is a more predictive indicator of team strength than is the team's win-loss record.

2.2 Generalization

It is not clear how James came up with his Pythagorean formula. A prolific baseball mind, he was known for combining his intuition for the game with a trial-and-error method for finding the formulas that would suit his purpose. The Pythagorean formula turned out to work quite well for predicting baseball team winning percentage. This leads to the natural question: *Is there a better formula for predicting team winning percentage?*

So that our notation applies more generally across sports, let us replace runs scored and runs allowed with points scored (PS) and points allowed (PA). There is an overwhelming set of possible functions that could relate points scored and allowed to winning percentage, but let us limit ourselves to those that take the same functional form as the Pythagorean formula:

$$\hat{\text{W\%}}^{\alpha}_{\text{Pythag}} = \frac{\text{PS}^\alpha}{\text{PS}^\alpha + \text{PA}^\alpha} = \frac{1}{1 + (\text{PS}/\text{PA})^{-\alpha}},$$

for some $\alpha \geq 0$. We call the ratio PS/PA the *score ratio*. If $\alpha = 0$, that corresponds to no relationship between score differential and winning percentage (because the Pythagorean winning percentage is identical to 50%). As α increases, so too does the strength of the relationship between score differential and winning percentage. As α tends to infinity, the Pythagorean winning percentage converges to 100% for any positive score differential (and 0% for any negative score differential).

To determine the best value for α, we introduce some mathematical notation to formalize the problem. Suppose we have a dataset of n teams, and for each team $i \in \{1, ..., n\}$, we observe the team's score ratio x_i and winning percentage $y_i \in [0, 1]$. Our objective is to find the value of α that leads to the best agreement between the observed winning percentages y_i and the Pythagorean winning percentags $1/(1 + x_i^{-\alpha})$. In mathematical terms, we wish to solve the following optimization problem:

$$\min_{\alpha} \left\{ \sum_{i=1}^{n} \left(y_i - \frac{1}{1 + x_i^{-\alpha}} \right)^2 \right\}.$$

This looks a little bit like a least squares regression problem, but the difference is that the expression being subtracted from y_i is not linear in α. So the formula for ordinary least squares does not apply in this case. One simple solution is to try many different values for α to find the one that leads to the lowest value of the objective function. With computing technology what it is today, we can get a precise estimate of the optimal α very quickly using this approach.

2.3 Consequences

The purpose of the Pythagorean formula is that it may be a more accurate predictor of future winning percentage than is the observed winning percentage. Winning percentage is a relatively coarse measure of team performance, whereas score ratio is a more precise measure of team performance. The way a team wins games is by scoring more points than the other team. The intuition is that perhaps all a team can control is its point ratio, and the team's winning percentage is just a noisy reflection of that point ratio. This statement is almost never literally true, but it is often approximately true.

One key consequence of the above is that if our goal is team evaluation or prediction of future game outcomes, we are generally better off basing our analysis on score differential rather than binary win/loss outcomes. Another important consequence is that if our goal is to measure player contributions to team success, we can often use scores as our unit of measurement. The objective of front office decision-makers is generally to maximize team winning (subject to budgetary constraints). Player evaluations based on scores are often less noisy than player evaluations based on wins, and the Pythagorean formula serves as a conversion between scoring and winning. For example, most implementations of baseball's Wins Above Replacement (WAR) are based first on a calculation of Runs Above Replacement.