WSABI Summer Research Lab **Summer 2025**

Lab 2: Simple Linear Regression

Instructor: Jonathan Pipping *Author: Ryan Brill*

2.1 Pythagorean Win Percentage

2.1.1 Data

We are given a dataset of team-seasons from 2017 to 2021. Each row i represents a team-season, and the columns are:

- i: the index of the i^{th} team-season.

- RS_i: the runs scored by the team over the course of the season.

- RA_i: the runs allowed by the team over the course of the season.

- WP_i: the win percentage of the team at the end of the season.

We want to predict end-of-season win percentage from Runs Scored and Runs Allowed. A team's deviation from this prediction is a measure of how **lucky** they were.

2.1.2 The Work of Bill James

Bill James, the godfather of sabermetrics (baseball analytics), created the Pythagorean Win Percentage formula.

$$\widehat{WP}^{(\text{Pythag})} = \frac{RS^2}{RS^2 + RA^2}$$

He made it up, and it works surprisingly well! We plot the data from the 2020 season in Figure 2.1.

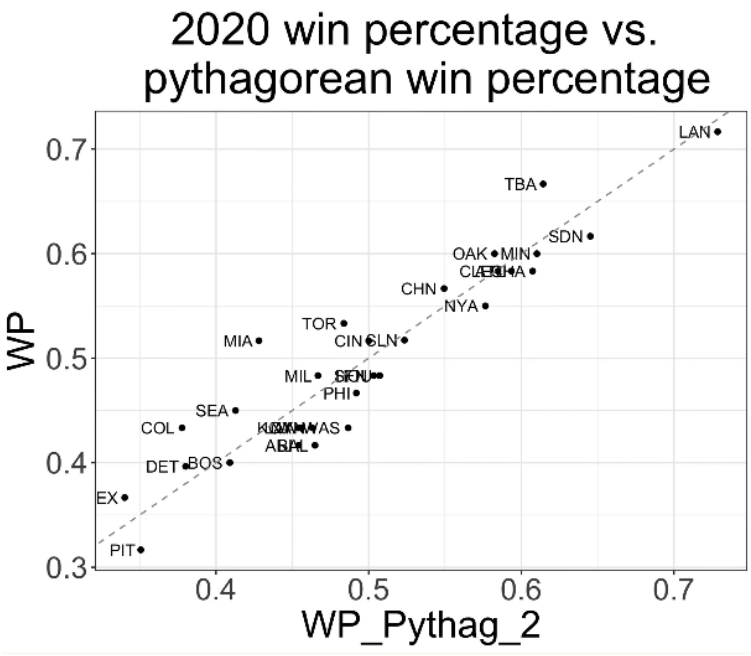

Figure 2.1: Win Percentage vs. Pythagorean Win Percentage for the 2020 season.

However, though the pythagorean coefficient (2) is quite good., it is arbitrary. You will improve on this using linear regression.

2.1.3 Your Assignment

Task 1: Use linear regression to find an exponent α so that the Pythagorean win percentage

$$\widehat{WP} = \frac{RS^\alpha}{RS^\alpha + RA^\alpha}$$

best fits the data.

Hint: You will first need to transform this equation to be **linear** in α. Try dividing both sides by RA^α.

Task 2: Create a visualization to show that $\widehat{WP}$ is better than $\widehat{WP}^{(\text{Pythag})}$.

2.2 Evaluating MLB General Managers

2.2.1 Data

We are given a dataset of MLB team payrolls and results for each season from 1998-2023. Each row i represents a team-season, and the columns are:

- i: the index of the i^{th} team-season.
- WP_i: the win percentage of the team at the end of the season.

- $\frac{\text{payroll}_i}{\text{median payroll}_i}$: the ratio of the team's payroll to the median payroll of all teams in the league that season.
- $\log(\frac{\text{payroll}_i}{\text{median payroll}_i})$: the log of the ratio of the team's payroll to the median payroll of all teams in the league that season.

We want to analyze the relationship between payroll and winning to evaluate the performance of MLB general managers.

2.2.2 Your Assignment

Task 1:

- Remove the 2020 Covid-shortened season.
- Plot winning percentage against payroll/median.
- Mark the Oakland A's and the New York Yankees' points on the plot.
- Add the regression line of WP as a function of payroll/median.
- Add the regression line of WP as a function of log(payroll/median).

Task 2:

- Now for each team-season, calculate the difference between the actual WP and the predicted WP using payroll/median.
- Do the same with log(payroll/median).
- Add these columns to the dataset.
- Find the average difference for each team and make a two graphs (one for each model) ordered from heighest to lowest.
- Change the y-axis scale to wins by multiplying by 162.
- Add a legend to the graphs.

2.2.3 Important Note:

Let $x = \frac{\text{payroll}}{\text{median payroll}}$ and consider the following models:

$$\text{Model A: } \widehat{WP}_i = \alpha_0 + \alpha_1 x_i$$
$$\text{Model B: } \widehat{WP}_i = \beta_0 + \beta_1 \log(x_i)$$

We interpret these models as follows:

- Model A: Increasing x_i by a constant value of 1 median payroll adds $\widehat{\alpha_1}$ to $\widehat{WP}_i$.

- Model B: Increasing x_i by $r \times 100\%$ adds $\widehat{\beta_1} r$ to $\widehat{WP_i}$.

Proof. Let $x_i' = (1+r)x_i$. Then,

$$\begin{aligned}\widehat{WP_i}' &= \beta_0 + \beta_1 \log(x_i') \\ &= \beta_0 + \beta_1 \log((1+r)x_i) \\ &= \beta_0 + \beta_1(\log(1+r) + \log(x_i))\end{aligned}$$

Since for small r, $\log(1+r) \approx r$, we have that

$$\begin{aligned}&= \beta_0 + \beta_1 \log(x_i) + \beta_1 r \\ &= \widehat{WP_i} + \beta_1 r\end{aligned}$$

□

Which model do you think is better intuitively? A or B?