Lab: Simple Linear Regression

1. Pythagorean Win Percentage

We are given a dataset of team-seasons from 2017 to 2021,

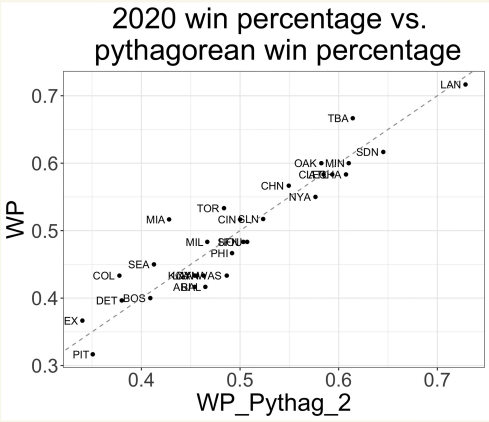
i = index of i^{th} team-season
RS_i = runs scored
RA_i = runs allowed
WP_i = win percentage

and 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 the team was.

Bill James, godfather of Sabermetrics (baseball analytics) and sports analytics, created Pythagorean Win percentage

$$\widehat{WP} = \frac{RS^2}{RS^2 + RA^2}$$

He made it up and it works quite well!



The pythagorean exponent is quite good, but is arbitrary.

- 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.

You'll need to transform this equation to be linear in α.

<u>Hint:</u> divide top and bottom by RS^{α}

- Create a simulation to show that

$$\widehat{WP} = \frac{RS^{\hat{\alpha}}}{RS^{\hat{\alpha}} + RA^{\hat{\alpha}}}$$ is better than $$\widehat{WP} = \frac{RS^2}{RS^2 + RA^2}.$$

2. Evaluating MLB general managers

We are given a dataset of MLB team payrolls and results for each season 1998–2023,

- row $i \leftarrow i^{th}$ team-season
- win percentage
- payroll/median payroll
- Log(payroll/median payroll)

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

(Try using Chat GPT Data Analyst for this question)

- Plot payroll/median against winning percentage, mark the Oakland A's and NY Yankees dots. Remove 2020.
  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).

- Now for each team-season calculate the difference between the actual WP and predicted WP using payroll/median and then log(payroll/median). Add this column to the dataset. Find the average difference for each team and make a graph ordered from highest to lowest. (one graph for each model).
Change the Y axis scale to win by multiplying by 162.