Lab: Simple Linear Regression

1. Pythagorean Win Percentage

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

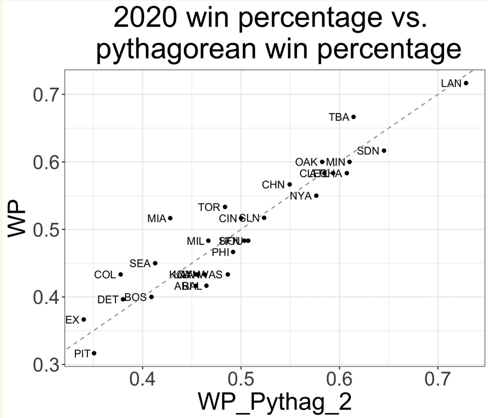
$$\begin{cases} i = \text{index of } i^{th} \text{ team-season} \\ RS_i = \text{runs scored} \\ RA_i = \text{runs allowed} \\ WP_i = \text{win percentage} \end{cases}$$

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

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

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.

__Hint:__ divide top and bottom by RS^α

- Create a visualization 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,

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

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

(try using ChatGPT 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 wins by multiplying by 162.

- <u>Note</u>:

$$\begin{cases} \text{Let } x = \text{payroll}/\text{median} \\ \text{Model A:} \quad WP = \alpha_0 + \alpha_1 \cdot x + \varepsilon \\ \text{Model B:} \quad WP = \beta_0 + \beta_1 \cdot \log(x) + \varepsilon \end{cases}$$

- Interpretation of Model A:

increasing x by a constant value of 1 median payroll adds $\hat{\alpha}_1$ to $\widehat{WP}$

- Interpretation of Model B:

increasing x by $r \times 100\%$ adds $\hat{\beta}_1$ to $\widehat{WP}$

<u>Proof</u> $x' = (1+r)x$ increase x by $r \cdot 100\%$

then

$$\begin{aligned} \widehat{WP}' &= \beta_0 + \beta_1 \log(x') \\ &= \beta_0 + \beta_1 \log[(1+r)x] \\ &= \beta_0 + \beta_1 \log(1+r) + \beta_1 \log(x) \\ &\approx \beta_0 + \beta_1 \log(x) + \beta_1 \cdot r \end{aligned}$$

since for r small, $\log(1+r) \approx r$

$$= \widehat{WP} + \beta_1 \cdot r$$

Which model is better intuitively?