

Assignment #1: The Big Game

Overview

In this assignment, you will leverage NFL game score data to predict the outcome of the Super Bowl. You will implement two distinct modeling philosophies: a discrete **Logistic Bradley-Terry** model based on wins and losses, and a continuous **Linear Bradley-Terry** model integrated with the **Pythagorean Expectation** formula.

Step-by-Step Instructions

Step 0: Data Acquisition

Download games from the current season using `nflreadr::load_schedules()`. You will need a table of all games from the 2025 season (including playoffs) with at least the following information: `game_id`, `away_team`, `away_score`, `home_team`, `home_score`.

Step 1: Empirical Calibration of Alpha (α)

Before modeling, determine the optimal exponent for the Pythagorean formula specific to the 2025 season. Define an error function that calculates the Sum of Squared Errors (SSE) between actual win percentages and those predicted by:

$$\text{Win \%} = \frac{\text{PF}^\alpha}{\text{PF}^\alpha + \text{PA}^\alpha}$$

Use `nls()` to find the α that minimizes SSE.

Step 2: Method 1: Logistic Bradley-Terry (Win/Loss)

Estimate a logistic Bradley-Terry model using `glm()` where the response is whether the home team won.

$$Z_i \sim \text{Bernoulli} \left(\frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)$$
$$\eta_i = \beta_0 + \beta_{h_i} - \beta_{a_i}$$

Calculate the head-to-head win probability for the Super Bowl matchup. How will you handle the fact that the Super Bowl is played at a neutral site?

Step 3: Method 2: Linear Bradley-Terry (Scoring)

Fit two separate linear Bradley-Terry models using the `lm()` function:

- **Model A:** Predict `score_diff = home_score - away_score`.
- **Model B:** Predict `score_total = home_score + away_score`.

Use these models to generate a predicted scoring margin (`score_diff`) and a predicted total score (`score_total`) for the Super Bowl matchup. Again, how will you handle the fact that the Super Bowl is played at a neutral site?

Step 4: Synthesis: Pythagorean Conversion

Convert the predicted margin and total from Step 4 into predicted home score and away score using the algebraic split:

$$\text{home_score} = \frac{\text{score_total} + \text{score_diff}}{2}, \quad \text{away_score} = \frac{\text{score_total} - \text{score_diff}}{2}$$

Finally, calculate the Super Bowl win probability using your **custom** α derived in Step 2.

Step 5: Table (PNG)

Using the `gt` package, produce a table comparing the Super Bowl finalists. The table must include:

- Win probability from Method 1 (logistic).
- Predicted home score and away score (Method 2).
- Win probability from Method 2 (Pythagorean).

Export this as `table.png`.

Step 6: Figure (PDF)

Create a scatter plot using `ggplot2` for all 32 NFL teams.

- **X-axis:** Linear BT Strength Coefficients (from Step 4, Model A).
- **Y-axis:** Logistic BT Strength Coefficients (from Step 3).
- **Features:** Highlight the Super Bowl finalists in a distinct color.

Export this as `figure.pdf`.

Submission Instructions

You must submit the following **three** files:

1. **Table (PNG):** The standalone `gt` table.
2. **Figure (PDF):** The standalone high-resolution scatter plot.
3. **R Code (R):** A `.R` script with your R code for completing the assignment.

Grading Rubric

1. Did you implement the Pythagorean formula correctly? (1pt)
2. Did you implement the logistic regression model correctly? (2pts)
3. Did you implement the linear regression models correctly? (2pts)
4. Did you calculate the Super Bowl predictions correctly? (1pt)

5. Did you produce the table correctly? (2pts)

6. Did you produce the figure correctly? (2pts)

Extra credit: 1pt for particularly well-presented figure and table.