WSABI Summer Research Lab **Summer 2025**

Lab 4: Logistic Regression

Instructor: Jonathan Pipping *Authors: RB, JP*

4.1 Field Goal Success Probability

4.1.1 Data

We have a dataset consisting of field goals, where each row represents a field goal. Each row includes the following variables:

- i: index of the i^{th} field goal
- y_i: 1 if the i^{th} field goal was made, and 0 otherwise
- ydl_i: the yardline of the i^{th} kick, measured in yards from the opponent's end zone.
- kq_i: kicker quality of the i^{th} kicker.
- kicker$_i$: the name of the kicker for the i^{th} field goal.

4.1.2 Your Task:

1. Model field goal success probability using at least 3 different models, with at least one being a linear regression and one being a logistic regression.
2. Use out-of-sample predictive performance to select the best model of the 3.
3. Write an interpretation of the coefficients of the selected model.
4. Plot the selected model's predictions against the actual outcomes.

4.2 Bradley-Terry NCAA Men's Basketball Power Scores

4.2.1 Data

We have a dataset consisting of the results of NCAA Men's Basketball games, where each row represents a game and includes the following variables:

- i: index of the i^{th} game
- s_i: season of the i^{th} game
- h_i: index of the home team
- a_i: index of the away team
- y_i: 1 if the home team won, and 0 otherwise

4.2.2 Bradley-Terry Model

The Bradley-Terry model supposes each team j has a latent power rating (or strength) β_j and the probability that team j beats team k at home is given by

$$p_{jk} = \text{Logistic}(\beta_0 + \beta_j - \beta_k)$$

and on the road is

$$p_{jk} = 1 - p_{kj} = \text{Logistic}(\beta_0 + \beta_k - \beta_j)$$

This is similar to the model we set up in Lecture 2, with the addition of the Logistic transformation. Note that β_0 is a home field advantage parameter.

4.2.3 Your Task:

1. Filter the data to include games from the 2023-2024 season (i.e. $s = 2023$)
2. Fit a Bradley-Terry model to the data.
3. Visualize the model coefficients and interpret them.
4. Set a Vegas spread for the 2023-2024 NCAA Tournament final between the Purdue Boilermakers and the UConn Huskies (Note: the game is played at a neutral site).

4.3 A Note on ELO Power Scores

ELO is an **"online"** or **rolling** version of Bradley-Terry logistic regression power scores, which is updated after every match. These models are frequently used in one-on-one sports, such as chess or tennis.

Let player i's ELO be β_i. Then the probability that player i beats player j is given by

$$\begin{aligned} p_{ij} = \mathbb{P}(i \text{ beats } j) &= \text{Logistic}(\beta_i - \beta_j) \\ &= \frac{1}{1 + \exp(\beta_i - \beta_j)} \end{aligned}$$

Then if player i beats j, update their ELO as follows:

$$\begin{aligned} \beta_i &\leftarrow \beta_i + K \cdot (1 - p_{ij}) \\ \beta_j &\leftarrow \beta_j + K \cdot (0 - p_{ij}) \end{aligned}$$

where K is a constant learning rate we set.