WSABI Summer Research Lab **Summer 2025**

Lab 3: Multivariable Linear Regression

Instructor: Jonathan Pipping *Author: Ryan Brill*

3.1 Four Factors in Basketball

Basketball has many different components that contribute to a team's success: shooting accuracy, defense, rebounding, etc. Which of these correlates with (or causes) winning? Which matters the most? Which matters the least?

3.1.1 The Work of Dean Oliver

Dean Oliver defined **four factors** that contribute to winning:

- Scoring: scoring accuracy
- Crashing: rebounding efficiency
- Protecting: turnover rates
- Attacking: free throw shooting

This was a big innovation, and when expressed in percentage terms, you can normalize for pace of play and get a consistent picture of what helps teams win.

3.1.2 Data

Your task will be to use multivariable regression to assess the relative impact of the four factors to winning. Let y be the team wins in a season (this is your outcome variable), and define the following four explanatory variables:

- $x_1 = \text{eFG\%} - \text{opp\_eFG\%}$
- $x_2 = \text{ORB\%} - \text{DREB\%}$
- $x_3 = \text{TOV\%} - \text{opp\_TOV\%}$
- $x_4 = \text{FT\_rate} - \text{opp\_FT\_rate}$

Here, eFG% is the effective field goal percentage, defined as

$$\text{eFG\%} = \frac{\text{FG} + \frac{1}{2}\text{3PT}}{\text{FGA}}$$

which weights successful three pointers 50% more than two pointers.

Also, TOV% is the turnover percentage, OREB% is the offensive rebounding percentage, DREB% is the defensive rebounding percentage, and FT_rate is the free throw rate: how often a team gets to the free throw line.

3.1.3 Your Assignment

Task 1: Get to know the data.

1. Find each variable's mean, standard deviation, maximum, and minimum.
2. Plot the marginal distributions of each variable.
3. Find the correlation between each pair of variables.

Task 2: Model the data.

1. Fit a multivariable linear regression model of the form

$$\mathbb{E}[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

2. Fit a second regression model after standardizing each of the x variables (to have mean 0 and standard deviation 1).
3. Which of these two models tells you about the relative value of each of the four factors to winning? Order them by importance to winning.
4. Which of the two models has better predictive performance and why? Use math or code an out-of-sample prediction test to show your answer.

3.2 Expected Outcome of a Punt

3.2.1 Data

We have a dataset consisting of punts, where each row represents a single punt. Each row contains the following variables:

- i: index of the i^{th} punt
- y_i: the outcome variable, which is the next yard line from the opponent's perspective after the punt.
- ydl_i: the yard line from which punt i took place (yards from opponent's goal line).
- pq_i: the punter quality of the punter who kicked punt i.
- punter$_i$: the name of the punter who kicked punt i.

3.2.2 Your Assignment

Task 1: Model the data.

1. Use multivariable regression to model the next yard line for the opponent as a function of current yard line and punter quality. Consider linear terms $\beta_1 ydl_i + \beta_2 pq_i$ and also transformations (e.g. quadratic, cubic, splines).

2. Select a model (e.g. by out-of-sample predictive performance) and visualize that model (e.g. plot expected next yard line vs current yard line for various punter quality values).

Task 2: Interpret the results.

1. Rank the punters by punt yards over expected (PYOE).
2. Visualize the rankings.