WSABI Summer Research Lab **Summer 2025**

Lab 6: Models Do What They're Told

Instructor: Jonathan Pipping *Authors: RB, JP*

6.1 Expected Points in American Football

6.1.1 What Are Expected Points?

Expected points added (EPA) is a popular tool for evaluating NFL team offenses, defenses, and quarterbacks. The EPA of a single play is calculated as the difference between the expected points (EP) before the play and the expected points after the play, or

$$\text{EPA}_{play} = \text{EP}_{\text{post-play}} - \text{EP}_{\text{pre-play}}$$

This calculation depends on how we define the expected points of a particular game state.

The expected points of a particular game state x attempts to capture the expected value of the **net points of the next score** in the half for the team possessing the ball. Expected points depends on many variables that comprise the game-state, such as yard line, down, yards to go for a first down, time remaining in the half, and some measure of relative team strength (e.g. the pre-game point spread between the two teams). For simplicity, we'll consider just these 5 covariates in our model.

6.1.2 Data

For this lab, we have play-by-play data from the 2019-2021 NFL seasons. Each row in the dataset corresponds to a single play and contains the following variables:

- `season`: the season where the play occurred.
- `game_id`: a unique identifier for the game.
- `play_id`: a unique identifier for each play within a game.
- `pts_next_score`: the net points of the next score in the half (relative to the team in possession).
- `label`: denotes visually which plays share the same "next score."
- `yardline_100`: the yard line, defined as the distance from the line of scrimmage to the opponent's end zone.
- `down`: the down of the play.
- `yds_to_go`: the yards to go for a first down.
- `half`: the half of the game.
- `half_seconds_remaining`: the number of seconds remaining in the half.
- `posteam_spread`: the pre-game point spread between the team in possession and the team not in possession (e.g. if the team in possession is favored by 3 points, `posteam_spread` is -3).

Let's use some of these variables to set up our model.

6.1.3 Setting Up Our Model

Let y be the net points of the next score in the half. For simplicity, we limit the possibilities for y to touchdowns, field goals, and safeties.

$$y \in \{-7, -3, -2, 0, 2, 3, 7\}$$

where positive values indicate a score for the team in possession and negative values indicate a score for the opposing team. A score of 0 indicates that the half ends without another score.

We then define the expected points of a game state x as

$$\mathbb{E}[y \mid x] = f(x)$$

where $f(x)$ is a function of the game state x. Our goal is to estimate $f(x)$, the conditional mean function, from the data. We've already learned a few ways to do this! For example, we could use a **multivariable linear regression**, which takes the form

$$y = x^T\boldsymbol{\beta} + \epsilon$$

where x is a vector of covariates for the game state, y is the net points of the next score in the half, and ϵ is a random error term. Then, the expected points of a game state x is given by

$$\text{EP}(x) = \mathbb{E}[y \mid x] = x^T\boldsymbol{\beta}$$

Anotheer way is a **multinomial logistic regression**, which takes the form

$$\log\left(\frac{\mathbb{P}(y = k \mid x)}{\mathbb{P}(y = k \mid x)}\right) = x^T\boldsymbol{\beta}_k$$

where $k \in \{-7, -3, -2, 2, 3, 7\}$. An equivalent model formulation is

$$\begin{cases} \mathbb{P}(y = k) = \frac{1}{1+\exp(-x^T\boldsymbol{\beta}_k)} \text{ for } k \neq 0 \\ \mathbb{P}(y = 0) = 1 - \sum_{k\neq 0} \mathbb{P}(y = k) \end{cases}$$

and then the expected points of a game state x is given by

$$\text{EP}(x) = \sum_k k \cdot \mathbb{P}(y = k \mid x)$$

Both of these models are additive, meaning they take the form

$$x^T\boldsymbol{\beta} = x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of the model.

6.1.4 Task 1: Build an Additive Model

Use the `multinom()` function from `R`'s `nnet` package to fit a multinomial logistic regression to model expected points. You will do this in a few ways:

1. Model expected points as a purely linear function of yard line.

$$\log\left(\frac{\mathbb{P}(y = k \mid x)}{\mathbb{P}(y = 0 \mid x)}\right) = \beta_{k0} + \text{yard line} \cdot \beta_{k1}, k \neq 0$$

   Plot your estimate of expected points (y axis) vs yard line (x axis). What's wrong with this model?

2. Use a spline on yard line to capture the non-linear relationship between expected points and yard line. Recall that you can use the `bs()` function from the `splines` package to do this. Plot your revised estimate of expected points (y axis) vs yard line (x axis).

3. Model expected points as a function of yard line and down. When you do this, be mindful of how you encode down (e.g. numeric vs categorical). Plot expected points (y axis) vs yard line (x axis) and color your points by down.

4. Model expected points as a function of yard line, down, and yards to go. When you do this, consider using a spline or log transformation! Plot expected points (y axis) vs yard line (x axis), coloring by yards to go and faceting by down.

5. Adjust your model by including time remaining in the half. Try both a linear term ($\beta *$time remaining) as well as a spline term. Plot expected points (y axis) vs yard line (x axis) on 1st down and 10 yards to go, coloring by time remaining.

What do we learn from this modeling process? **Models do what they're told**, capturing (only) the trends that we tell them to capture.

6.1.5 A Problem with Expected Points

Public expected points models essentially end here and do not adjust for team quality! They argue that there is no need to do this, since for team/player evaluation we are only interested in the expected points of an **average offense facing an average defense**. Is this a valid justification?

6.1.6 Task 2: Adjust for Team Quality

1. Let's call the best expected points model from Task 1 M. Now, make model M' which adjusts for pre-game point spread (e.g., using a linear term).

2. Plot the expected points from M' with point spread $= 0$ vs the expected points from M. Are these functions the same? Why or why not? Justify your answer with a visualization.

6.2 A Note on Selection Bias

What we observe in expected points models is **selection bias**, a pervasive issue in sports analytics and many other fields that use observational (or non-experimental) data. Note that *good teams have more plays in our dataset than bad teams*, and *good teams score more points than bad teams*. What does this really mean? Observed data **over-samples** good teams, meaning the expected points for the average of the observed teams differs from the true expected points for average teams!

6.2.1 Discussion Question

With this in mind, are the following quantities the same or different and why?

- The percentage of all 3-point attempts made in the NBA this year.

- The "true" 3-point make percentage of an average NBA player.

If they're different, which would we expect to be higher? How could we adjust for this?