CAUTION: These lecture notes are under construction. You may find parts that are incomplete.

4 Plus-Minus Models

In the previous chapter, we learned about using point differentials to estimate team strengths. For team sports, we are often interested in estimating the strengths of individual players. This is particularly important for teams who have to make player personnel decisions in the form of trades, free agent signings and draft picks. In this chapter we will learn about one approach to player evaluation based on point differential *when the player is on the field*. The stat *Plus-Minus*, popularized in hockey and basketball, is just that—the score differential when the player is active in the game.

To formalize this mathematically, we will introduce some notation. We define a *stint* to be a period of time during which no substitutions occur, meaning that the active players do not change during a stint. Assume we have a dataset of stints numbered $i = 1, 2, ..., n$ (which may span multiple games) and players numbered $j = 1, 2, ..., p$. For each stint i, we observe the following:

- w_i: the stint length, which may be measured by time (e.g. hockey) or possessions (e.g. basketball);
- y_i: the *normalized* score differential (per unit of stint length) in favor of the home team; and
- x_{ij}: the +1/–1/0 indicator that player j is active for home/away/neither team in stint i.

The purpose of introducing the normalized score differential is to make apples-to-apples comparisons across stints of differenti lengths. For player j, their cumulative Plus-Minus is $\sum_i w_i x_{ij} y_i$, and their average Plus-Minus is:

$$s_j = \frac{\sum_{i=1}^{n} w_i \cdot x_{ij} \cdot y_i}{\sum_{i=1}^{n} w_i \cdot x_{ij}^2}. \tag{1}$$

DISCUSSION: What are strengths and weaknesses of using Plus-Minus for player evaluation?

4.1 Adjusted Plus-Minus

One drawback of Plus-Minus is that if one player typically plays alongside a particularly strong teammate, this will inflate their Plus-Minus. In other words, Plus-Minus does not control for quality of competition *or* quality of teammates. Our method for addressing this is directly analogous to the strength-of-schedule adjustment we learned using the Bradley-Terry model in the previous chapter. We use the random variable Y_i to represent the score differential of stint i, and we model the distribution of Y_i as follows:

$$\begin{aligned} \eta_i &= \beta_0 + \sum_{j \in H_i} \beta_j - \sum_{j' \in A_i} \beta_{j'} \\ Y_i &\sim \text{Normal}(\eta_i, \sigma^2 / w_i). \end{aligned} \tag{2}$$

We have $p+1$ regression coefficients to estimate: one β for each player (interpretable as the player's strength), as well as β_0 (interpretable as home-field advantage). As in the previous chapter, this model is not identifiable, so we must introduce an additional constraint. The simplest, most common constraint is to set $\beta_1 = 0$, meaning that the first player is the reference player against which all other players are measured. Having established this constraint, we proceed with p regression coefficients to estimate.

Note one key difference between equation (2) and the Bradley-Terry model from the previous chapter. In the Bradley-Terry model, the variance of Y_i was assumed to be σ^2 for each i. In the Adjusted Plus-Minus model, the variance of Y_i is σ^2 / w_i, which is different for each i. This reflects the intuition that there is more random noise involved in the normalized score differential for shorter stints. Because of this, when

estimating the model, we want to put more weight on minimizing the error for longer stints. In contrast with OLS, our criterion for estimating the regression coefficient vector $\boldsymbol{\beta}$ is *weighted* least squares (WLS):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} w_i(y_i - (\beta_0 + \beta_{h_i} - \beta_{a_i}))^2 \qquad \text{s.t. } \beta_1 = 0. \tag{3}$$

Note that this choice of $\hat{\boldsymbol{\beta}}$ is the *maximum likelihood estimator* of the model specified by equation (2).

We now introduce matrix notation to calculate the WLS estimator of the regression coefficients. We use $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ to denote the $n \times 1$ vector of score differentials; we use $\mathbf{X}$ to denote the $n \times p$ sparse matrix of regression covariates; and we use $\mathbf{W}$ to denote the $n \times n$ diagonal matrix of observation weights:

$$(\mathbf{X})_{ij} = \begin{cases} +1 & \text{if } j = 1 \text{ (intercept column)} \\ x_{ij} & \text{otherwise} \end{cases}, \qquad (\mathbf{W})_{ii'} = \begin{cases} w_i & \text{if } i = i' \\ 0 & \text{otherwise} \end{cases}.$$

Lastly, we use $\boldsymbol{\beta} = (\beta_0, \beta_2, \beta_3, ..., \beta_p)^T$ to denote the $p \times 1$ vector of regression coefficients. Then the WLS estimate of $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}.$$

We interpet $\hat{\beta}_j$ as the estimated strength of player j. As with the Bradley-Terry model, the Adjusted Plus-Minus model comes with the following satisfying property:

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} w_i \left(x_{ij} \cdot y_i - \sum_{j' \neq j} \hat{\beta}_{j'} x_{ij'} \right)}{\sum_{i=1}^{n} w_i \cdot x_{ij}^2}.$$

Compare this expression with the definition of average Plus-Minus in equation (1). Observe that the estimated strength of player j is equivalent to their average Plus-Minus, after adjusting for the estimated strengths of all other players involved in player j's stints. Hence the name: *Adjusted* Plus-Minus.

> Discussion: What are strengths and weaknesses of using Adjusted Plus-Minus for player evaluation?

4.2 The Rasch Model

So far we have seen two models (Bradley-Terry and Adjusted Plus-Minus) that are variations on the same concept. Whereas the Bradley-Terry model estimates team strengths, the Adjusted Plus-Minus model estimates player strengths. One may think of the Bradley-Terry model as a special case of a the Adjusted Plus-Minus model, where every matchup involves one home player and one away player (and every stint is the same length). In this section we introduce one more model which is yet another variation on this same core concept.

One commonality between Bradley-Terry and Adjusted Plus-Minus is that each team (or player) can appear on either side of each matchup: home or away. The strength of the team (or player) is the same regardless of the side on which they appear (although there is an effect for home-field advantage). For many adversarial interactions in sports, this restriction is not appropriate. Consider, for example, the matchup between a batter and a pitcher in baseball. While it is true that batters may pitch and that pitchers may bat on occasion, it is not reasonable to assume that a player's batting strength is equal to their pitching strength. Enter the Rasch model.

The Rasch model comes from the field of psychometrics. Originally it was used to model the performance of students on test questions. Each student is assumed to have an ability, and each question is assumed to have a difficulty. The probability of a successful answer is a function of the sum of the student's ability and the question's difficulty. This framework applies well to many sports applications. For example, we may assume that each batter has an ability and each pitcher has a difficulty, or vice versa.

For simplicity, we will describe the Rasch model in the context of modeling game scores. Unlike the Bradley-Terry model, however, we will estimate separate offensive and defensive strengths for each team. Assume we observe a set of team-scores y_i (two team-scores per game) numbered $i = 1, 2, ..., n$ involving

teams numbered $j = 1, 2, ..., p$. For each team-score i, o_i is the team that did the scoring, and d_i is the team that allowed the scoring. Using Y_i to represent the random variable for team-score i, we model the distribution of Y_i as follows:

$$\eta_i = \beta_0 + \beta^O_{o_i} + \beta^D_{d_i}$$
$$Y_i \sim \text{Normal}(\eta_i, \sigma^2).$$

Here we have introduced superscripts on β^O to represent offensive strength and β^D to represent defensive strength (more negative = stronger defense). We have $2p + 1$ regression coefficients to estimate: one β^O for each team; one β^D for each team; and the intercept β_0. Once again, this model is not identifiable until we introduce the additional constraint $\beta^O_1 = \beta^D_1 = 0$.

To fit the Rasch model, we must estimate the $(2p-1)\times 1$ vector $\boldsymbol{\beta} = (\beta_0, \beta^O_2, \beta^O_3, ..., \beta^O_p, \beta^D_2, \beta^D_3, ..., \beta^D_p)^T$. We are back in the land of OLS, so we can use the familiar formula once we first establish the necessary matrix notation. As before, the $n \times 1$ vector $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ contains the observed team-scores. We construct our $n \times (2p-1)$ matrix $\mathbf{X}$ by concatenating $\mathbf{1} = (1, 1, ..., 1)^T \in \mathbb{R}^n$, $\mathbf{X}^O \in \mathbb{R}^{n\times(p-1)}$, and $\mathbf{X}^D \in \mathbb{R}^{n\times(p-1)}$:

$$(\mathbf{X}^O)_{ij} = \begin{cases} 1 & \text{if } o_i = j-1 \\ 0 & \text{otherwise} \end{cases}, \qquad (\mathbf{X}^D)_{ij} = \begin{cases} 1 & \text{if } d_i = j-1 \\ 0 & \text{otherwise} \end{cases}, \qquad \mathbf{X} = (\mathbf{1}, \mathbf{X}^O, \mathbf{X}^D).$$

Note that in this case, $\mathbf{X}$ is a sparse matrix with at most three nonzero entries (all equal to one) per row. With this established, the OLS estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

The Rasch model is a generalization of the Bradley-Terry model in that the actors involved (teams or players) can have separate roles and separate strengths in each of these roles. We could generalize this further by allowing for non-normally distributed outcomes (as we saw with the Bradley-Terry model for win-loss outcomes) and/or by allowing for unequally weighted observations (as we saw with the Adjusted Plus-Minus model). Yet another way to generalize the Rasch model further is to allow for more than two roles. For example, one may hypothesize that the catcher and umpire could have some effect on the outcome of a baseball pitch (especially the ball/strike call if the batter does not swing). In this case, the linear term in the model would take the form $\eta_i = \beta_0 + \beta^B_{b_i} + \beta^P_{p_i} + \beta^C_{c_i} + \beta^U_{u_i}$, including effects from batter, pitcher, catcher and umpire. This is a flexible model that can describe the probability distribution of many different outcomes of interest in sports.

Discussion: How is the Rasch model similar to the Bradley-Terry model? How is it different?

4.3 Quiz Questions

1. Why do we use Weighted Least Squares (WLS) instead of Ordinary Least Squares (OLS) to estimate the Adjusted Plus-Minus model?

2. Which is it: Is the Adjusted Plus-Minus model a special case of the Bradley-Terry model? Or is the Bradley-Terry model a special case of the Adjusted Plus-Minus model?