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

2 Base-Out Run Expectancy and Linear Weights

In Chapter ??, we learned where wins come from: Wins are composed of runs. But where do runs come from? That is the focus of this chapter.

2.1 Markov Chain Model

Baseball is often described as a sport that lends itself particularly well to statistical analysis. The primary reason is that a baseball game is composed of discrete events. First, one batter faces one pitcher, resulting in an outcome. Then, a second batter comes to the plate and produces a new outcome. And so on. This makes it relatively straightforward to isolate the impact of individual players on the number of runs scored by either team. The first building block of this analysis is the *base-out run expectancy*: Given the bases occupied and the number of outs, what is the expected number of runs that will score in the remainder of the inning?

To define base-out run expectancy, we start with the *Markov chain* model. A Markov chain is a probability model consisting of a set $\mathcal{S}$ of states and a transition probability function $p : \mathcal{S} \times \mathcal{S} \to [0, 1]$ between the states. We observe a sequence of states, and the probability of transitioning from one state to the next depends only on the current state. When using a Markov chain to model data, how we define the state is important modeling decision. We want the state to include all of the information necessary for determining the probabilities of transitioning to each possible subsequent state, and at the same time we prefer a simpler, more parsimonious model.

In baseball, the most common application of the Markov chain model is to describe the progression of an inning as a sequence of static states between plate appearances. We define the *base-out state* to be (b_1, b_2, b_3, o), where $b_k \in \{0, 1\}$ indicates whether base k is occupied, for $k \in \{1, 2, 3\}$; and $o \in \{0, 1, 2\}$ represents the number of outs at the beginning of a plate appearance. Every inning starts in state $(0, 0, 0, 0)$. In addition to the 24 ($= 2\times2\times2\times3$) non-terminal states, we need five terminal states (r) for $r \in \{0, 1, 2, 3, 4\}$ (corresponding to the number of runs scored on the final transition—necessary for calculations below).

With the state defined, what remains is to define the transition probabilities between states. One could approach this different ways, but the most common approach is to use the empirical transition probabilites observed in a chosen dataset. For example, if we observe the state $(0, 0, 1, 0)$ 100 times in our dataset, and 60 of those times the next state is $(0, 0, 1, 1)$, then our estimated transition probability from $(0, 0, 1, 0)$ to $(0, 0, 1, 1)$ is 60%. Because we are often working with big samples of data (the typical MLB regular season has approximately 170,000 plate appearances), these empirical transition probabilities are generally reasonable estimates. We will use $p(s, s')$ to denote the probability of transitioning from state s to state s'.

2.2 Base-Out Run Expectancy

Using the Markov chain model for the progression of an inning, we can calculate the expected number of runs scored from any base-out state to the end of the inning. We use $r(s, s')$ to denote the reward (i.e. the number of runs scored) on the transition from state s to state s'. We can write it as follows:

$$r(s, s') = \begin{cases} (b_1 + b_2 + b_3 + o) + 1 - (b'_1 + b'_2 + b'_3 + o') & \text{if } s' \text{ is not terminal} \\ r' & \text{if } s' \text{ is terminal} \end{cases}$$

We use $v(s)$ to denote the value (i.e. the rest-of-inning run expectancy) of state s. The value function satisfies the following recursive relationship (a simplified version of the Bellman equation):

$$v(s) = \sum_{s' \in \mathcal{S}} p(s, s')\{r(s, s') + v(s')\}$$

To calculate $v(\cdot)$, we initialize $v(s) = 0$ for all $s \in \mathcal{S}$ and then iterate the above equation until convergence.

2.3 Player Evaluation

2.3.1 RE24

We come to our first player evaluation metric of the course. From $v(\cdot)$ we have the run expectancy of each base-out state. RE24 is the change in run expectancy averaged across a batter's plate appearances:

$$\text{RE24}(b) = \frac{\sum_{i=1}^{n} \mathbb{I}\{b_i = b\}(r_i + v(s_i') - v(s_i))}{\sum_{i=1}^{n} \mathbb{I}\{b_i = b\}}.$$

2.3.2 Linear Weights

To calculate linear weights, we start with a similar calculation to RE24, but we average the change in run expectancy within outcome, rather than within batter. The linear weight of outcome o is given by:

$$\ell(o) = \frac{\sum_{i=1}^{n} \mathbb{I}\{o_i = o\}(r_i + v(s_i') - v(s_i))}{\sum_{i=1}^{n} \mathbb{I}\{o_i = o\}}.$$

Once we have the linear weight $\ell(\cdot)$ of each outcome, the metric LW is simply the average of these linear weights across a batter's plate appearances:

$$\text{LW}(b) = \frac{\sum_{i=1}^{n} \mathbb{I}\{b_i = b\}\ell(o_i)}{\sum_{i=1}^{n} \mathbb{I}\{b_i = b\}}.$$

Discussion: What are the advantages and disadvantages of RE24 and LW relative to each other?

2.4 Regression to the Mean

There is an analogy to be drawn between the RE24/LW relationship and the relationship between winning percentage and Pythagorean record from the previous chapter. Just like winning percentage, RE24 is measurement that carries more descriptive meaning (it measures what actually matters). Just like Pythagorean record, LW is a more stable measurement. We saw in the previous chapter that Pythagorean record is a better predictor of future winning percentage than winning percentage itself (unless the sample size is more than several hundred games). One might ask a similar question for RE24 and LW: When do we switch to preferring RE24 over LW?

Let's instead acknowledge that the question presents a false dichotomy. A better question is: *How can we use both RE24 and LW to best predict future RE24?* This is where regression to the mean comes in. Recycling notation from the previous chapter, for batter $j \in \{1, ..., p\}$, we use n_j to denote the number of plate appearances, and we use the random variable Z_j to represent the residual $RE24(j) - LW(j)$.

$$\begin{aligned} Z_j &\sim \text{Normal}(\eta_j,\, \sigma_Z^2/n_j) \\ \eta_j &\sim \text{Normal}(0,\, \sigma_\eta^2). \end{aligned}$$

Recognizing this as a Bayesian model, we can use Bayes' rule to derive the posterior distribution of η_j given $Z_j = z_j$:

$$\eta_j \mid Z_j = z_j \sim \text{Normal}\left(\frac{n_j/\sigma_Z^2 \cdot z_j}{n_j/\sigma_Z^2 + 1/\sigma_\eta^2}, \frac{1}{n_j/\sigma_z^2 + 1/\sigma_\eta^2}\right).$$

In the previous chapter, we chose between ignoring the residual or fully including it in our prediction:

$$X_j = X_j + 0 \qquad \text{vs.} \qquad Y_j = X_j + Z_j$$

Now we have derived a third option that is a weighted average of the two extremes:

$$X_j + \frac{n_j/\sigma_Z^2 \cdot Z_j}{n_j/\sigma_Z^2 + 1/\sigma_\eta^2}.$$

As opposed to making a sudden switch from one extreme to the other, this third estimator smoothly transitions from ignoring the residual to giving it full weight. We make the following observations:

1. When $n_j = \sigma_Z^2/\sigma_\eta^2$ (the point at which our preference between X_j and Y_j flips in the previous chapter), our estimator is $X_j = Z_j/2$.

2. As $n_j \to \infty$, our estimator converges to $Y_j = X_j + Z_j$.

3. As $n_j \to 0$, our estimator converges to X_j.