

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 BASE-OUT RUN EXPECTANCY

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 mathematically, we start with the *Markov chain* model. A Markov chain is a probability model consisting of states and probabilities of transitioning between each pair of 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. Let's define the *base-out state* to be (b_1, b_2, b_3, o) , where $b_i \in \{0, 1\}$ indicates whether base i is occupied, for $i \in \{1, 2, 3\}$; and $o \in \{0, 1, 2, 3\}$ represents the number of outs at the beginning of a plate appearance. The only state for which $o = 3$ is $(0, 0, 0, 3)$, the terminal state (end of inning). Thus there are a total of 25 $(= 2 \times 2 \times 2 \times 3 + 1)$ base-out states: one for each combination of bases occupied and outs (for outs less than three), plus the terminal state. Every inning starts in state $(0, 0, 0, 0)$.

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 probabilities 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' , and we will use P to denote the 25-by-25 matrix of transition probabilities.

Given this formulation of the Markov chain model, we can leverage properties of the Markov chain to calculate base-out run expectancy.

2.2 RE24

2.3 LINEAR WEIGHTS