A MARKOV CHAIN MODEL FOR RUN PRODUCTION IN BASEBALL

Thomas W. Polaski

Winthrop University

February 12, 2013

Introdcution: Baseball by the Numbers

Baseball and numbers seem to go together. Statistics have been kept for over a hundred years, but lately sabermetrics has taken this obsession to a new level. Mathematical models of baseball are used to compare eras, players, teams, and even stadia. Three aspects of baseball make mathematical models tractable:

- relatively small number of configurations
- relatively small number of events that can happen
- discrete nature

Possible Base Configurations and Events

Bases Occupied:

None
1st only, 2nd only, 3rd only,
1st and 2nd, 1st and 3rd, 2nd and 3rd,
1st, 2nd and 3rd

Outs:

0, 1, 2, or 3

Events:

single, double, triple, home run, walk, hit batsman, out, ...

Markov Chains

A **Markov chain** is a mathematical model for movement between **states**. A process starts in one of these states and moves from state to state. The moves between states are called **steps** or **transitions**. The chain can be said to move between states and to be "at a state" or "in a state" after a certain number of steps.

The state of the chain at any given step is not known, but the **probability** that the chain is at a given state after n steps depends only on

- the state of the chain after $n-1$ steps, and
- the probabilities that the chain moves from one state j to another state i in one step.

The Transition Matrix P

These probabilities are called **transition probabilities** for the Markov chain. The transition probabilities are placed in a matrix called the **transition matrix** P for the chain by entering the probability of a transition from state j to state i at the (i, j)-entry of P. So if there were m states named 1, 2, ... m, the transition matrix would be the $m \times m$ matrix

$$P = \begin{array}{c} \text{From:} \\ \begin{array}{cccc} 1 & \text{j} & & \text{m} \end{array} \\ \begin{bmatrix} & \vdots & & \\ & \downarrow & & \\ & p_{ij} & \cdots & \rightarrow \\ & & & \end{bmatrix} \end{array} \begin{array}{c} \text{To:} \\ 1 \\ \\ \\ \text{i} \\ \text{m} \end{array}$$

The Big Idea

Use a Markov chain to model run production in baseball.

- States could be different possible configurations.
- Transition probabilities could be calculated from actual data.

States of the Markov Chain

Transient States

Bases Occupied	Outs	Bases Occupied	Outs
None	0	1st,2nd	1
1st	0	1st,3rd	1
2nd	0	2nd,3rd	1
3rd	0	1st,2nd,3rd	1
1st,2nd	0	None	2
1st,3rd	0	1st	2
2nd,3rd	0	2nd	2
1st,2nd,3rd	0	3rd	2
None	1	1st,2nd	2
1st	1	1st,3rd	2
2nd	1	2nd,3rd	2
3rd	1	1st,2nd,3rd	2

Absorbing States

Left on Base	Outs	Left on Base	Outs
0	3	2	3
1	3	3	3

Transition Probabilities

- Only the following events are allowed for the player at bat: single, double, triple, home run, walk, hit batsman, out.
- No base stealing, errors, sacrifices, sacrifice flies, double plays, triple plays, or other ever-more-bizarre events are allowed.
- For the following analysis to work, each batter must have the same probability of singling, doubling, etc. Each batter is identical.

Outcomes of Batting Events

Batting Event	Outcome
Walk or Hit Batsman	The batter advances to first base. A runner on first base advances to second base. A runner on second base advances to third base only if first base was also occupied. A runner on third base scores only if first base and second base were also occupied.
Single	The batter advances to first base. A runner on first base advances to second base. A runner on third base scores. A runner on second advances to third base with probability $1 - p$ and scores with probability p. ($p = 0.63$)
Double	The batter advances to second base. A runner on first base advances to third base. A runner on second base scores. A runner on third base scores.

Outcomes of Batting Events (cont.)

Triple	The batter advances to third base. A runner on first base scores. A runner on second base scores. A runner on third base scores.
Home Run	The batter scores. A runner on first base scores. A runner on second base scores. A runner on third base scores.
Out	No runners advance. The number of outs increases by one.

Calculation of Transition Probabilities

The transition probabilities are calculated from the above rules and from aggregate data. For example, the data from the 2012 MLB season is:

Batting Event	Occurrences	Probability
Walk or Hit Batsman	16203	0.089
Single	27941	0.154
Double	8261	0.046
Triple	927	0.005
Home Run	4934	0.027
Out	123128	0.679

Example: The probability of a transition from "runner on first, 0 out" to "runners on first and second, 0 out" is the probability of a walk or a hit batsman or a single, which is $0.089 + 0.154 = 0.263$.

The Transition Matrix P

The transition probabilities are calculated as above and placed in the transition matrix P. We order the states listing absorbing first, then transient.

- The probability of a transition from an absorbing state to itself is 1.
- The probability of a transition from an absorbing state to any other state is 0.
- The probabilty of a transition from a transient state is governed by the above probabilities.

$$P = \left[\begin{array}{c|c} I & S \\ \hline O & Q \end{array}\right]$$

The Fundamental Matrix M

The **fundamental matrix** for the Markov chain is $M = (I - Q)^{-1}$. Suppose that the chain starts at state i. Then

- The expected number of visits to state j before being absorbed is the (j, i)-element in M.
- The expected number of steps until absorption is the sum of the (j, i)-elements of M, where j ranges over all transient states.

In terms of baseball, the sum of the i^{th} column of M is the expected number of batters that will come to bat during the remainder of the inning starting from state i.

The sum of the first column of M is the expected number of batters that will bat in an inning. We call this value $E(B)$.

The Matrix $A = SM$

The matrix $A = SM$ is important to our analysis. Suppose that the chain starts at the transient state i. Then

- The (j, i)-element of the matrix $A = SM$ is the probability that the Markov chain is eventually absorbed at state j.

In the baseball model, the i^{th} column of A contains the probabilities that 0, 1, 2, or 3 runners are left on base at the end of the half-inning starting at state i.

Thus the first column of A can be used to compute the expected number of runners left on base:

$$0 \cdot P(0 \textit{ left}) + 1 \cdot P(1 \textit{ left}) + 2 \cdot P(2 \textit{ left}) + 3 \cdot P(3 \textit{ left})$$

This value is called $E(L)$.

The Baseball Conservation Law

Every batter who comes to the plate makes an out, scores, or is left on base.

Let B be the number of batters who come to the plate in an inning, R be the number of runs scored, and L be the number left on base.

$$B = 3 + R + L, \text{ or } R = B - L - 3$$

Taking expected values,

$$E(R) = E(B) - E(L) - 3$$

and we can calculate $E(B)$ and $E(L)$, thus $E(R)$.

Does this Model Work?

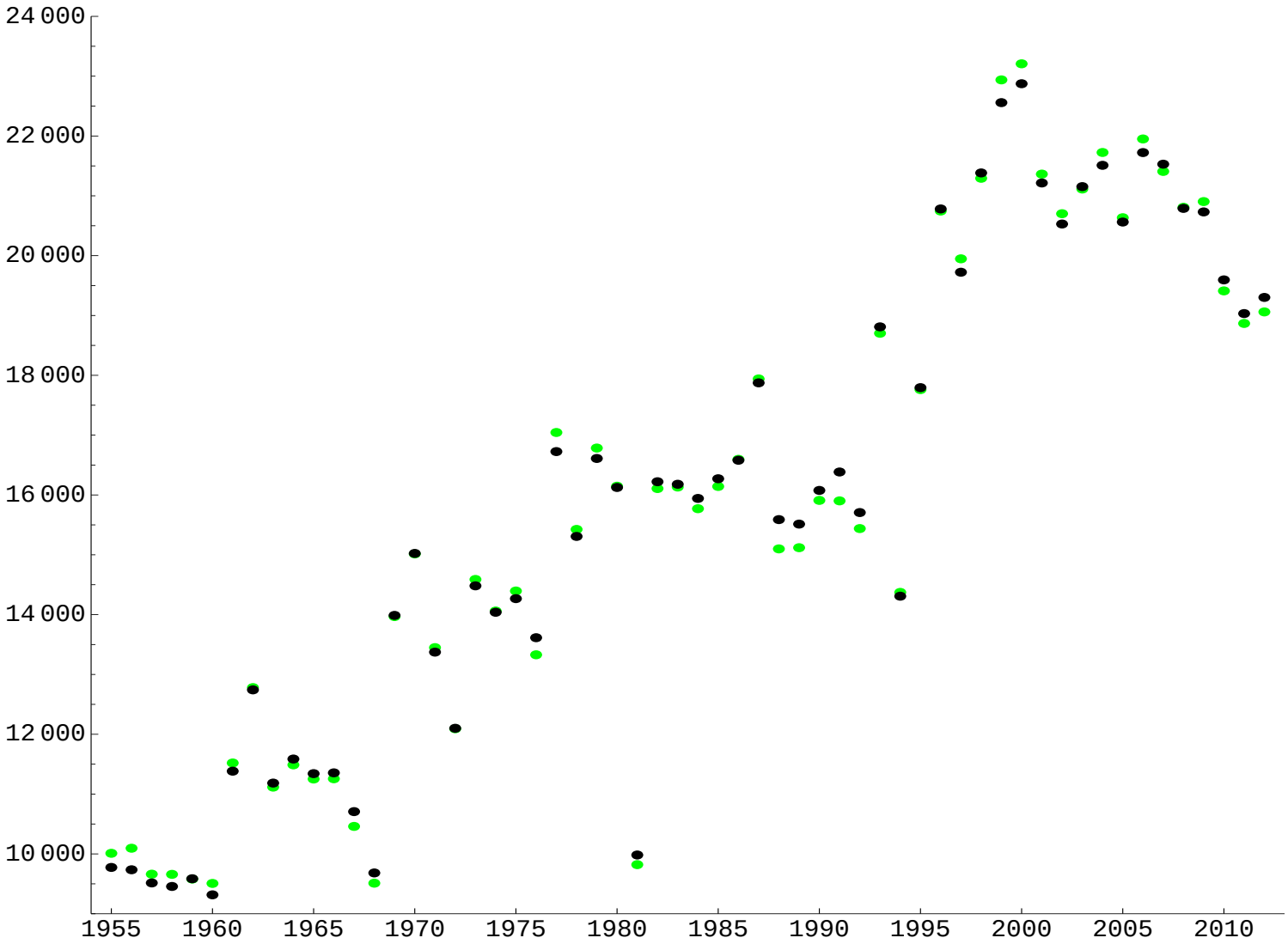
For the 2012 MLB season, the model predicts that 0.439618 earned runs will score in a single inning on the average. Since 43,355 innings were played during the season, the model predicts that 19,059.6 earned runs would score during the season.

In reality, 19,302 earned runs scored during the season, or 0.445208 per (half)-inning. The error in the model was -242.4 earned runs for the season; the relative error was -1.26%.

Does this Model Work?

Year	Earned Runs Scored	Model Prediction	Error	Relative Error
2000	22874	23207.0	333.0	1.46%
2001	21215	21363.2	148.2	0.70%
2002	20529	20701.2	172.2	0.84%
2003	21154	21114.0	-40.0	-0.19%
2004	21510	21724.7	214.7	1.00%
2005	20562	20634.3	72.34	0.35%
2006	21723	21950.6	227.6	1.05%
2007	21529	21407.8	-121.2	-0.56%
2008	20790	20808.4	18.4	0.09%
2009	20731	20904.0	173.0	0.83%
2010	19595	19411.1	-183.9	-0.94%
2011	19032	18866.9	-165.1	-0.87%
2012	19302	19059.6	-242.4	-1.26%
Total	270546	271152.9	606.9	0.22%

Earned Runs Scored (●) and Model Predictions (●) 1955-2012



Even More Results

In a similar manner, we can also compute expected runs scoring after any state has been reached. For the 2012 MLB season we get:

Bases Occupied	Outs	Expected Runs	Bases Occupied	Outs	Expected Runs
None	0	0.4400	1st,2nd	1	0.8609
1st	0	0.7817	1st,3rd	1	0.9135
2nd	0	0.9959	2nd,3rd	1	1.0746
3rd	0	1.0285	1st,2nd,3rd	1	1.4002
1st,2nd	0	1.3317	None	2	0.0913
1st,3rd	0	1.3852	1st	2	0.1894
2nd,3rd	0	1.5593	2nd	2	0.3008
1st,2nd,3rd	0	2.0189	3rd	2	0.3464
None	1	0.2384	1st,2nd	2	0.4113
1st	1	0.4565	1st,3rd	2	0.4500
2nd	1	0.6177	2nd,3rd	2	0.5614
3rd	1	0.6843	1st,2nd,3rd	2	0.7325

Applications

- Player Rating
  - Consider the same player batting over and over.
  - Use the player's batting statistics to generate transition probabilities.
  - Run model to find the expected number of earned runs the player would score in a single inning or a nine-inning game.
- Strategy
  - Compare the expected number of runs starting from certain states to decide on whether to sacrifice or steal.
  - Example: Sacrifice with man on first and no outs.