Prediction of Future Performance in Baseball

Shane T. Jensen
Blake McShane
Adi Wyner
ESPN-Wharton Collaboration

• Ongoing collaboration between our sports group and ESPN for the evaluation of MLB players
• The overall goal is a comprehensive evaluation of pitching, hitting and fielding performance
• Primary attention in industry is on retrospective performance in past seasons
• We are more interested in prospective performance in future seasons
  • Should you be giving out contracts to players for past or future performance?
Projection of Hitting Performance

• We want to predict future hitting performance of individual players, accounting for:
  • Age, Position, Ball-park Differences

• We also want our predictions to balance the player’s personal history with the history of the population of players at his position
  • Need this balance to be a function of the length of player’s personal history and his consistency
  • Need to still make intelligent predictions for players with very little personal history

• Bayesian framework provides principled way to achieve this balance
Available Data

- Our focus is predicting rates of major hitting events $G = \{\text{BB/PA, IBB/PA, 1B/AB, 2B/AB, 3B/AB, HR/AB}\}$

- For each player $X$ and hitting event $G$, we have their entire **personal history** at the season-level
  - Past hitting rates: $Y_1, \ldots, Y_n$
  - Past ages: $\text{age}_1, \ldots, \text{age}_n$
  - Past home ballparks: $\text{park}_1, \ldots, \text{park}_n$
  - Past positions: $\text{pos}_1, \ldots, \text{pos}_n$

- We use this data for all baseball players (1970-2004) to estimate our **population models**
Step 1: Position-Specific Trajectory

- Hitting rate $Y$ modeled as quadratic function of age
  - Allows for both increasing and decreasing periods of ability over career

$$Y_i = \beta_{0p} + \beta_1 \cdot \text{park}_i + \beta_2 \cdot \text{age}_i + \beta_3 \cdot \text{age}_i^2 + \varepsilon_i$$

- Coefficients which control the age trajectory are different for each position $p$

- Model includes park effect $\beta_1$ which allows us to estimate influence of each different ball-park on $Y_i$

- $\varepsilon_i$ represents a player’s departure from the age trajectory in season $i$
  - Need projected departure $\varepsilon_{n+1}$ to calculate projected $Y_{n+1}$
Estimated Trajectory Curves for HR
Estimated Trajectory Curves for 3B
Accounting for Personal History

- Age trajectories are estimated from the entire population and do not take into account player’s personal history.
  - Projections for players with long and consistent careers should not be as influenced by the rest of population.

- Solution: create a projected departure from age trajectory that is a compromise between their personal history and the population at their position.

\[ \text{Projected } \varepsilon_{n+1} = w \cdot \text{Past History} + (1-w) \cdot \text{Population} \]

- Need a principled function \( w \) that factors in length and consistency of personal history for each player.
Population as a Prior Distribution

• Player X’s **personal history** is observed past departures from age trajectory: \( \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \)

  Observed Data: \( \varepsilon_i \sim \text{Normal} (\theta, \sigma^2) \)

• Parameter \( \theta \) is the player’s true departure from age trajectory, which is unknown with **prior distribution**:

  Prior Distribution: \( \theta \sim \text{Normal} (\mu_p, \tau_p^2) \)

• Prior parameters are based on the **population of past players** at Player X’s position \( p \)
Bayes Rule

- Bayes rule combines personal history (likelihood) with population (prior) into posterior distribution:

\[
\theta \sim \text{Normal}\left(\frac{\tau_p^2}{\tau_p^2 + \sigma^2/n} \cdot \bar{\epsilon} + \frac{\sigma^2/n}{\tau_p^2 + \sigma^2/n} \cdot \mu_p, \frac{\tau_p^2 \cdot \sigma^2/n}{\tau_p^2 + \sigma^2/n}\right)
\]

- We use posterior mean as our projected departure:

\[
\mathcal{E}_{n+1} = \frac{\tau_p^2}{\tau_p^2 + \sigma^2/n} \cdot \bar{\mathcal{E}} + \frac{\sigma^2/n}{\tau_p^2 + \sigma^2/n} \cdot \mu_p
\]

- Empirical Bayes: use estimates of variances \(\sigma^2\) and \(\tau^2\)
Weighting Function

• The weight $w$ for each player is a function of the length and variance of their career and their position.

$$W = \frac{\tau_p^2}{\tau_p^2 + \sigma^2 / n}$$

<table>
<thead>
<tr>
<th>Player</th>
<th>HR Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Delgado</td>
<td>0.99</td>
</tr>
<tr>
<td>M. Ramirez</td>
<td>0.99</td>
</tr>
<tr>
<td>C. Beltran</td>
<td>0.88</td>
</tr>
<tr>
<td>R. Cano</td>
<td>0.52</td>
</tr>
</tbody>
</table>

- Long, consistent career
- Long but inconsistent career
- Short career
Examining our Population Prior

- For several positions and rates (eg. shortstops and HR/AB), the population distribution of departures from age trajectory seems to contain a mixture:
  - Most SS are not good HR hitters, but a subset seem to show consistently higher departures.
  - Perhaps don’t want to shrink these power-hitting SS all the way down to overall SS mean.
Improvement 1: Mixture Model for Prior

• Expand our prior population model to be a mixture of **standard** players and **elite** players

• **Elite probability** for a player is determined by personal history relative to other players at the same position
  • Each position may have a different proportion of elite players for each hitting event

<table>
<thead>
<tr>
<th>Elite HR Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Rodriguez 3B 0.99</td>
</tr>
<tr>
<td>D. Jeter SS 0.93</td>
</tr>
<tr>
<td>D. Willis P 1.00</td>
</tr>
<tr>
<td>I. Suzuki RF 0.00</td>
</tr>
<tr>
<td>S. Finley CF 0.59</td>
</tr>
</tbody>
</table>

• Elite probability used to calculate prior departure $\mu^*$ for a player as mixture of **elite** and **standard** pop. means
  • more refined prior “guess” at the appropriate population for a particular player
Improvement 2: Temporal Weighting

- When summarizing a player’s past performance, we want to give more emphasis to recent history.
- Instead of using simple averaging for elite probabilities and past mean departures, use weighted average with more weight towards more recent seasons:

![Linear Weight](chart1.png) ![Exponential Weight](chart2.png)
Putting it all together

For each player and each hitting rate $G$:

1. Estimate players hitting rate next year: $X_{n+1}$ from trajectory curve

   \[ X_{n+1} = \beta_0 + \beta_1 \cdot \text{park}_{n+1} + \beta_2 \cdot \text{age}_{n+1} + \beta_3 \cdot \text{age}^2_{n+1} \]

2. Calculate player’s elite status and use to produce refined prior estimate of departure $\mu^*$ from population model

3. Calculate projected departure as a compromise of prior estimate and observed data (past personal departures):

   \[ \varepsilon_{n+1} = w \cdot \overline{\varepsilon} + (1 - w) \cdot \mu^* \]

4. Combine departure estimate and trajectory curve for a final projection for hitting rate $G$:

   \[ Y_{n+1} = X_{n+1} + \varepsilon_{n+1} \]
Variance of Projections

- Modeling approach allows us to also estimate the **variance** of each player’s projection
- Variance of each projection is also a function of length and consistency of player’s career

<table>
<thead>
<tr>
<th>Long and consistent career</th>
<th>Short but consistent career</th>
<th>Long but inconsistent career</th>
<th>Short and inconsistent career</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Delgado</td>
<td>M. Ramirez</td>
<td>K. Griffey Jr.</td>
<td>S. Taguchi</td>
</tr>
<tr>
<td>0.008</td>
<td>0.010</td>
<td>0.020</td>
<td>0.026</td>
</tr>
</tbody>
</table>
Prediction Intervals

• For each player, we combine their projection and their variance into a **prediction interval**

• Validated 2005 prediction intervals (using exponential temporal weighting) against observed rates for 2005:

<table>
<thead>
<tr>
<th>Event</th>
<th>Coverage Percentage</th>
<th>Average Predicted Rate</th>
<th>Average Width of Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singles</td>
<td>83</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>Doubles</td>
<td>82</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>Triples</td>
<td>80</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Home Runs</td>
<td>82</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Walks</td>
<td>83</td>
<td>0.08</td>
<td>0.11</td>
</tr>
</tbody>
</table>

• Results are slightly worse for rarer events (e.g. triples)
Model Validation

• Compare observed 2005 rates to our model predictions with different internal settings:
  • Prior: none (no shrinkage) vs. mixture vs. no mixture
  • Temporal weighting: none vs. linear vs. exponential

• Also compare to two external alternatives:
  • Dumb alternative: 2004 observations carried forward
  • Smart alternative: PECOTA - industry standard

• PECOTA uses detailed historical matching plus manual adjustment - hard to beat!
## Validation Results for HR/AB rate

<table>
<thead>
<tr>
<th>Model</th>
<th>Population Prior</th>
<th>Temporal Prior</th>
<th>% increase/decrease relative to PECOTA Bias</th>
<th>SD</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PECOTA</td>
<td>None</td>
<td>None</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2004 Carried Forward</td>
<td>None</td>
<td>None</td>
<td>109</td>
<td>133</td>
<td>123</td>
</tr>
<tr>
<td>Our Model without Shrinkage</td>
<td>None</td>
<td>None</td>
<td>57</td>
<td>119</td>
<td>115</td>
</tr>
<tr>
<td>Our Model with Shrinkage</td>
<td>Mixture</td>
<td>None</td>
<td>75</td>
<td>103</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>Linear</td>
<td>93</td>
<td>108</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>Mixture</td>
<td>Exponential</td>
<td>93</td>
<td>108</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>No Mixture</td>
<td>Exponential</td>
<td>60</td>
<td>107</td>
<td>107</td>
</tr>
</tbody>
</table>

- 2004 carried forward is awful, PECOTA is hard to beat
- Compared to PECOTA, our model reduces bias but not lower RMSE
- Best bias reduction from model without shrinkage, but this model also has higher SD (and RMSE)
- No-mixture prior seems to do better than mixture prior
- HR/AB rate is the worst rate for our model: some of the other rate predictions beat PECOTA predictions
Summary and Future Improvements

• Bayesian model provides principled balance between a player’s personal history and overall population
  • Comparable quality between our automated method and manually-curated PECOTA
  • Our shrinkage is position-specific: position is used as a proxy for other unmeasured covariates

• Need better park effects: trying to separate our actual effect of park from effect of the team

• Need to also model number of plate appearances (PA) in order to convert rates into totals
  • PA totals are highly variable between seasons due to injury and utility vs. regular players
  • Hidden Markov Model: latent injury/utility states
  • Can also use HMM to improve model for elite status
Acknowledgements

• Projection project is collaboration with Adi Wyner and Blake McShane at Wharton and Ted Mann at ESPN

• Data comes from The Baseball Archive run by Sean Lahman

• Other members of ESPN-Wharton sports group: Kenny Shirley, Michael Frieman, Elan Fuld, Matt Carruth, Matt Koizim