Batter Pitch Mix Breakdown-Technical Report

Patrick Mellady

2024-10-19

Objective

We are interested in a solution to the following question:

Given pitch-by-pitch data from the 2021-2023 MLB regular seasons, we would like you to predict the pitch mixes that batters faced in 2024 across three groups: fastballs, breaking balls, and off-speed pitches

In this report, we shall outline the theory behind our model, the model's implementation in R, and the shortcomings of our model.

Theoretical Development

The chosen model will be a parametric probability model. To implement this model, we will make the assumption that each pitch a batter sees is a multinomial random variable with three response categories - fastball, breaking ball, and off-speed - where we have allocated each pitch type in the data to one of the three aforementioned categories according to their definitions to Baseball Savant. Thus, our distributional assumption is

$$Y_i \sim MN_3(1, \pi_{i1}, \pi_{i2}, \pi_{i3})$$

The problem is to estimate π_{i1}, π_{i2}, and π_{i3} given a vector of covariates, x_i, for each observation. We do this using a multicategory logit model of the form

$$\ln(\frac{\pi_{ir}}{1 - \sum_{j \neq r} \pi_{ij}}) = x_i^T \beta_r$$

We have copious amounts of data available for our analysis, so we will need to perform some variable selection on our covariates. We begin by only considering data we would know *prior* to a pitch being thrown, since this is the same information a pitcher would have before throwing any particular pitch. This is information such as the batter, the count, the position of runners on the base paths, and so on. Further, to perform the analysis at the season level, we don't want to predict the mix of pitches at the pitch-by-pitch level. So, we will transform some of these variables into indicators of certain game states and aggregate these for each year. For example the we will use the binary covariate COUNT to represent a hitter's count, or RISP to represent that there are runners in scoring position. We will then sum those indicators over the course of a whole season, giving an aggregated count of how many times a hitter appears in particular game states within a year.

The aggregated counts serve a secondary purpose: they give us information on the frequency each game state is seen by each hitter. Using this information on the game states, we can project how many of each game state will be seen by each hitter in 2024. This will give us the necessary covariates to predict the pitch mix breakdown for the 2024 season

We will train our model on the data from the years 2021 through 2023 and predict the covariate values for 2024 by taking the median frequency of each game state from the previous years. After fitting the model and predicting the covariates, we will predict the pitch distribution for each player in 2024.

Implementation

The model will be fit in R using the glmnet package. Typically, the glmnet package is used to add regularization to the covariates, however, is incredibly optimized and can be modified to not penalize covariates. The speed of the glmnet function, as well as it's compatibility with the multicategory logit model, works extremely well with the large design matrix in the problem.

Limitations and Advantages

The limitations of this model also present some potential advantages. Since we have aggregated the data within a year, we lose information about particular pitcher/hitter match-ups as well as how teams approach pitching to certain hitters. This limits the applicability of the current model as a tool for preparing for specific match-ups. However, the general form of the model allows for us to tweak this limitation whenever desired, meaning that we can change the aggregation technique to be for certain pitchers, teams, or game states and see a expected hitters pitch mix breakdown in more specific settings. This allows for the model to be applied in a practice setting or for in-game plate appearance preparation.

Performance

To evaluate the performance of the model, we will randomly sample 5 hitters (using the sample() function in R) and compare the projected 2024 pitch mixes to the actual pitch mixes seen during the 2024 season. The model predicts

2024 Projected Pitch Mixes

PLAYER_NAME	PITCH_TYPE_FB	PITCH_TYPE_BB	PITCH_TYPE_OS
Pham, Tommy	0.565	0.327	0.108
Díaz, Elias	0.541	0.335	0.124
Bell, Josh	0.527	0.245	0.228
Turner, Trea	0.528	0.351	0.120
O'Neill, Tyler	0.540	0.361	0.099

The corresponding 2024 pitch mixes for those players is

2024 Observed Pitch Mixes

PLAYER_NAME	PITCH_TYPE_FB	PITCH_TYPE_BB	PITCH_TYPE_OS
Pham, Tommy	0.559	0.335	0.106
Díaz, Elias	0.542	0.336	0.122
Bell, Josh	0.555	0.221	0.224
Turner, Trea	0.504	0.366	0.131
O'Neill, Tyler	0.543	0.333	0.124

which gives a total mean square error of

```
## [1] 0.003826
```