Modeling the Probability of a Strikeout for a Batter/Pitcher Matchup

Glenn Healey, *Fellow, IEEE*

Abstract—We analyze models for predicting the probability of a strikeout for a batter/pitcher matchup in baseball using player descriptors that can be estimated accurately from small samples. We start with the log5 model which has been used extensively for describing matchups in sports. Log5 is a special case of a logit model and we use constrained logistic regression over nearly one million matchup observations to assess the use of the log5 explanatory variables for this application. We also show that a batter/pitcher ground ball rate interaction variable is significant for the prediction of strikeout probability and we provide physical justification for the inclusion of this variable in the model. We quantify the differences among the models and show that batters control the majority of the variance in predicted strikeout rate.

Index Terms—Modeling, prediction, sports analytics, baseball, log5

1 Introduction

In recent years, a large amount of data has been collected that allows for the detailed characterization of the performance of baseball players. The ability to use this data to predict the distribution of outcomes for a batter/pitcher matchup has important implications for team building, player usage, and forecasting systems. Before a season begins, for example, a general manager might tune his roster to optimize his team's expected performance against the distribution of starting pitchers in his league or division [11]. A field manger could use a prediction system during the season to optimize his lineup according to the opponent's starting pitcher on a given day. Predictions could also be used to inform in-game decisions such as which relief pitcher to deploy against a particular sequence of batters in the opposing lineup.

Ideally, a large sample of previous outcomes for a particular batter/pitcher matchup would be available to support prediction, but sufficient samples are rarely available for this purpose [6], [18], [19]. An approach that allows building larger samples [17], [19] is to partition batters and pitchers into groups where the members of each group have similar characteristics. This process allows study of the distribution of outcomes when a batter from a particular group faces a pitcher from a particular group and has led to important discoveries about the dependence of matchup outcomes on the group membership of the batter and pitcher [19]. Researchers have shown, however, that the effectiveness of using a player's performance against a group to predict the outcome of an individual batter/pitcher confrontation is limited by sample sizes that are still insufficient [19] and by sensitivity to the location of group boundaries in the feature space [17].

In 1983, James [8] presented the log5 method that he developed with Adams for predicting the probability of an outcome in a binary experiment that matches two players. The log5 formula (which is also known as the James function) accounts for the individual success rates of the players as well as the average success rate for the environment. The method has been shown [7] to have a number of desirable and interesting mathematical properties and, by accounting for the environment, can support predictions over a range of outcomes and contexts. Baseball researchers [5], [13] have analyzed large sets of data and concluded, for example, that the method accurately predicts the probability of a hit in a batter/pitcher matchup. The approach has also been used to predict the probability of other outcomes [2] such as whether a batter reaches base in a confrontation with a pitcher.

We will examine the properties of log5 and related models after first partitioning matchups into one of four platoon configurations (LHP versus LHB, LHP versus RHB, RHP versus LHB, RHP versus RHB) according to the handedness (left or right) of the batter and pitcher. The platoon configuration has a significant impact on the expected outcome of a matchup [19] and since handedness is a discrete variable this partitioning does not require a process for defining group boundaries. Within each platoon configuration, we follow the spirit of the log5 model by representing each player using descriptors that are derived using nearly all of their matchups for a given season and by representing contextual information using league averages. Extending this study to include more flexible models for sharing information across players may provide a useful direction for future work. Latent variable models, for example, have recently been exploited for spatial modeling in basketball [14], [20].

We will consider the particular problem of modeling the probability of a strikeout using all major league play-by-play data from 2003 to 2013. Strikeouts are a useful starting point for this analysis since they are largely independent of factors outside the control of the batter and pitcher such as team

• *The author is with the Department of Electrical Engineering and Computer Science, University of California, Irvine, CA 92617. E-mail: ghealey@uci.edu.*

Manuscript received 29 Sept. 2014; revised 26 Jan. 2015; accepted 19 Mar. 2015. Date of publication 25 Mar. 2015; date of current version 3 Aug. 2015. Recommended for acceptance by A. Banerjee.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2416735

defense which can affect the probability of a hit or ballpark dimensions which can affect the probability of a home run. HITf/x measurements [9] extend play-by-play data by quantifying quality of contact using estimates of the speed and direction of batted balls. This data would be useful for generating more detailed matchup models that include predictions for batted balls, but HITf/x data is not publicly available at this time. The approach taken in this paper, however, can be generalized as additional data becomes available.

Batters and pitchers will be represented by strikeout rate and ground ball rate which have been shown to reach a high level of reliability for sample sizes that are achievable for individual players within a platoon configuration using data for a single season [3], [4]. We begin by using constrained logistic regression to investigate a model for expected strikeout rate that uses the same explanatory variables as log5. This work builds on the data modeling efforts of Staude [15], [16] for predicting strikeout rates. We further show that a variable that accounts for the interaction between batter and pitcher ground ball rates is significant in predicting the probability of a strikeout for a matchup. We analyze the differences between the various models and provide examples for cases where the models differ. We also use the models to show that batters control a larger fraction of the variance in expected strikeout rate than pitchers.

2 Matchup Modeling

2.1 Log5 Model

In this section, we describe the use of the log5 model [8] for the specific problem of predicting the probability of a strikeout for a plate appearance that matches a batter and a pitcher. Let K be a random variable where $K = 1$ if a matchup ends in a strikeout and $K = 0$ otherwise. We can represent the probability of a strikeout for a given matchup using

$$P(K = 1|\mathbf{x}) = P(K = 1|x_1, x_2, \ldots, x_n), \tag{1}$$

where $\mathbf{x}$ is a vector of n explanatory variables. For the special case of the log5 model, we let B be a batter's strikeout rate and let P be a pitcher's strikeout rate in a league with a strikeout rate of L. The log5 method then predicts the probability E^* of a strikeout for a matchup between a batter and a pitcher according to

$$E^* = \frac{(BP)/L}{(BP)/L + (1-B)(1-P)/(1-L)}. \tag{2}$$

This equation has several interesting properties. E^* remains the same if B and P are interchanged and $E^* = L$ if the pitcher and batter are both league average ($B = P = L$). In addition, E^* predicts that a batter will have a strikeout rate above B when facing a pitcher with an above average strikeout rate ($P > L$) and will have a strikeout rate below B when facing a pitcher with a below average strikeout rate ($P < L$). Similar statements also hold for pitchers due to the symmetry of the log5 equation. If we denote the odds ratios for B, P, L, and E^* by

$$B_o = \frac{B}{1-B}, \quad P_o = \frac{P}{1-P}, \quad L_o = \frac{L}{1-L}, \quad E^*_o = \frac{E^*}{1-E^*} \tag{3}$$

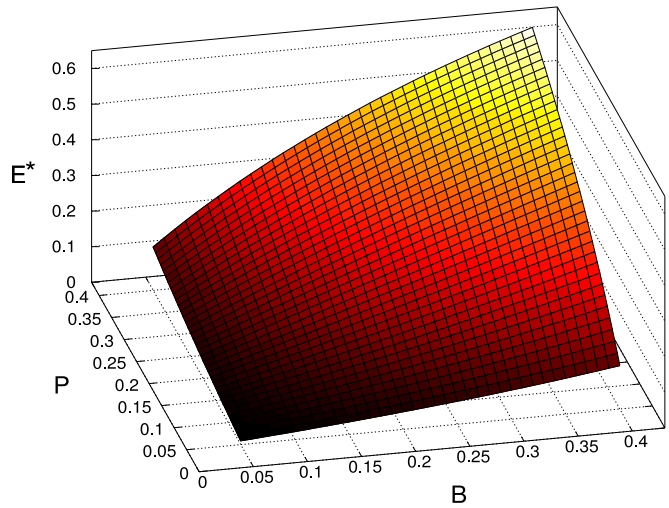


Fig. 1. Log5 predicted strikeout probability (RHP versus RHB, 2013).

then the log5 formula of Equation (2) can be written more compactly as

$$E^*_o = (B_o P_o)/L_o \tag{4}$$

so that

$$\ln(E^*_o) = \ln(B_o) + \ln(P_o) - \ln(L_o). \tag{5}$$

As an example, if we set L to the league strikeout rate for 2013 matchups between right-handed pitchers and right-handed batters, then the log5 model predicts the surface $E^*(B, P)$ shown in Fig. 1.

2.2 Logit Model

For many applications, the general model of Equation (1) can be represented using

$$E = P(K = 1|\mathbf{x}) = P(K = 1|x_1, x_2, \ldots, x_n) = F(S), \tag{6}$$

where S is a constant c plus a linear combination of the explanatory variables

$$S = c + c_1 x_1 + c_2 x_2 + \cdots + c_n x_n \tag{7}$$

and $0 < F(S) < 1$ for all real numbers S so that the probability E is between zero and one. A common choice for $F(S)$ is the logistic function which is defined by

$$F(S) = \frac{1}{1 + e^{-S}}. \tag{8}$$

The odds ratio of E in (6) is then given by

$$E_o = \frac{F(S)}{1 - F(S)} = e^S \tag{9}$$

which leads to

$$\ln(E_o) = S = c + c_1 x_1 + c_2 x_2 + \cdots + c_n x_n. \tag{10}$$

By comparing Equations (5) and (10), we see that the log5 model E^* is equivalent to the logit model E defined by Equations (6)-(10) if $n = 2, x_1 = \ln(B_o), x_2 = \ln(P_o)$ and the constants are given by $c = -\ln(L_o), c_1 = 1.0$, and $c_2 = 1.0$.

3 Model Assessment

3.1 Explanatory Variables

In Section 2.2 we showed that the log5 model is a special case of a logit model. This leads to several questions. We can ask whether the logit model can improve on the log5 model if the same explanatory variables ($x_1 = \ln(B_o)$ and $x_2 = \ln(P_o)$) are used but the c, c_1, and c_2 constants are assigned different values. We can also ask if the use of additional explanatory variables can improve the accuracy of the predicted strikeout rate. These questions can be answered by evaluating the performance of different models against sets of data.

3.1.1 Descriptor Reliability

Before models can be evaluated, we need to determine a set of candidate player descriptor variables that can be used to define the model explanatory variables. An important consideration is the size of the sample that is required to estimate a descriptor variable reliably. Let X_i be the observed value for a candidate descriptor variable, e.g. strikeout rate, for player i over a sample of N plate appearances. Then we can write

$$X_i = T_i + E_i, \tag{11}$$

where the observed X_i is an estimate of the player's true ability T_i and E_i is random error. The size of the random error decreases as the sample size N increases, but using a sample that extends across multiple years brings the possibility of significant changes in the true skill level (T_i) of the player within the sample. Ideally, we would like to estimate a separate set of descriptors for each applicable platoon configuration for each player for each year. This necessitates the use of descriptors that can be estimated reliably from samples that are consistent with this requirement.

Given a set of either batters or pitchers and N plate appearance observations for each player in the set, the reliability [10] for a player descriptor variable is defined by

$$R = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}, \tag{12}$$

where σ_X^2 is the variance of the observed X_i over the players, σ_T^2 is the variance of the true ability T_i over the players, and σ_E^2 is the variance of the random error. An increase in the sample size N leads to an increase in R which corresponds to a reduction in the fraction of the total variance σ_X^2 that is due to the variance σ_E^2 of the random error.

Carleton [3], [4] used the Kuder-Richardson [12] method for estimating reliability to determine the minimum sample size N that is required to achieve a reliability of 0.7 for a number of player descriptor variables. He showed that, for both batters and pitchers, strikeout rate and ground ball rate are the variables that reach $R = 0.7$ at the smallest N. In particular, Carleton's studies found that strikeout rates reach a reliability of 0.7 after 60 plate appearances for batters and after 70 plate appearances for pitchers. Thus, strikeout rate can be estimated with a high reliability within a platoon configuration for many individual batters and pitchers using a single season of data. He also showed that a reliability of 0.7 is achieved for ground ball rate after 80 balls in play for batters and after 70 balls in play for pitchers. We note that balls in play do not occur during every plate appearance as events such as strikeouts and walks are not considered balls in play. In 2013, for example, approximately 71.4 percent of plate appearances resulted in a ball in play which means that, on average, the threshold values of 70 and 80 balls in play for pitchers and batters are achieved after 98 and 112 plate appearances respectively. Thus, ground ball rates require more plate appearances to reach a reliability of 0.7 than strikeout rates, but can also be estimated with a high reliability within a platoon configuration for many individual batters and pitchers using a single season of data. The other variables that Carleton considered required significantly larger samples to reach a reliability of 0.7 and would, therefore, provide less reliable descriptors for a player when considering data for a single season and platoon configuration.

TABLE 1
Player Descriptors for Left-Handed Batter David Ortiz and Right-Handed Pitcher Max Scherzer for 2013

Player_ID	Role	Year	Configuration	SO Rate	GB Rate
ortid001	Batter	2013	RHP versus LHB	0.150000	0.392593
ortid001	Batter	2013	LHP versus LHB	0.160377	0.414634
schem001	Pitcher	2013	RHP versus RHB	0.315476	0.373206
schem001	Pitcher	2013	RHP versus LHB	0.269777	0.378882

3.1.2 Descriptor Computation

Based on the discussion in Section 3.1.1, we represent each batter and pitcher by their strikeout rate and ground ball rate within each applicable platoon configuration for each season. Before computing these rates to characterize each player and the league overall, we remove all plate appearances that ended with a bunt or an intentional walk. We also remove all plate appearances where a pitcher was batting. We refer to plate appearance totals after bunts, intentional walks, and pitchers batting have been removed as adjusted plate appearances. Strikeout rate is defined as the ratio of strikeouts to adjusted plate appearances and ground ball rate is defined as the ratio of ground balls to balls in play. As an example, Table 1 gives the individual player descriptors for left-handed batter David Ortiz and right-handed pitcher Max Scherzer for the 2013 season.

The strikeout and ground ball rate for batters, pitchers, and the league can change from year to year and can also have a significant dependence on the platoon configuration. Fig. 2, for example, plots the league average strikeout rate L and Fig. 3 plots the league average ground ball rate L_G for each of the four platoon configurations for the years from 2003 to 2013. The rates clearly depend on the platoon configuration and can change from year to year. We see, for example, that same-sided matchups (LHP versus LHB and RHP versus RHB) lead to higher strikeout rates and that strikeout rates in general have been rising steadily over the years considered in Fig. 2.

3.2 Regression Analysis

Our model evaluation process uses play-by-play data provided by Retrosheet. Accurate batted ball data which

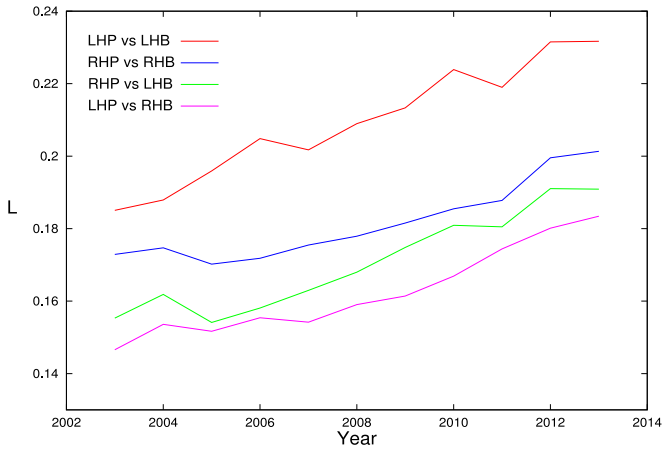


Fig. 2. League average strikeout rate.

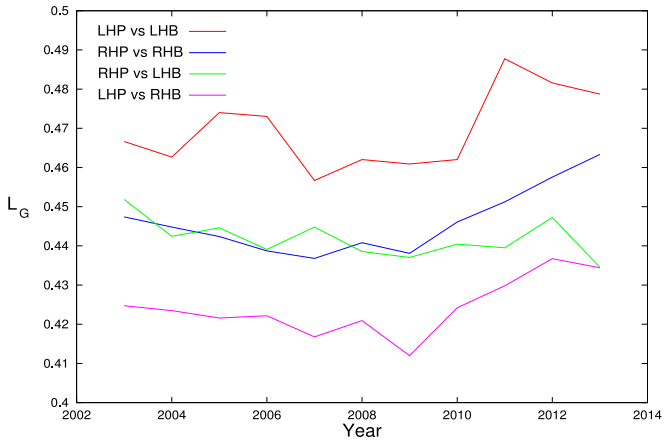


Fig. 3. League average ground ball rate.

TABLE 2
Number of Observations Used for Each Platoon Configuration Over the Years 2003 to 2013

LHP versus LHB	LHP versus RHB	RHP versus LHB	RHP versus RHB
24,419	124,992	405,861	441,000

is required to estimate ground ball rates has only been available since the 2003 season. For model evaluation, therefore, we consider all plate appearances except bunts, intentional walks, and pitchers as batters from 2003 to 2013 that involve both a batter and pitcher for which reliable rates against the handedness of the opponent can be computed. For the reasons discussed in Section 3.1.1, we use 150 adjusted plate appearances against the handedness of the opponent in the season under consideration as a threshold for both the batter and pitcher for a plate appearance to be considered. Table 2 gives the number of plate appearance observations from 2003 to 2013 that pass this adjusted plate appearance threshold for each platoon configuration. We note that there are significantly fewer observations for matchups that involve left-handed pitchers.

Given the set of observations associated with a platoon configuration, logistic regression can be used to estimate the coefficients for the logit model described in Section 2.2

TABLE 3
Binary Logit Output, LHP Versus LHB, 24,419 Observations

Variable	Coefficient	Std. Error	z-Statistic	p-value
$\ln(L_o)$	−1.001638	0.064578	−0.025361	0.9798
$\ln(B_o)$	0.970395	0.041492	−0.713520	0.4755
$\ln(P_o)$	1.031243	0.049611	0.629755	0.5289

TABLE 4
Binary Logit Output, LHP Versus RHB, 124,992 Observations

Variable	Coefficient	Std. Error	z-Statistic	p-value
$\ln(L_o)$	−0.946265	0.029952	1.794048	0.0728
$\ln(B_o)$	0.987263	0.020049	−0.635292	0.5252
$\ln(P_o)$	0.959001	0.021781	−1.882338	0.0598

TABLE 5
Binary Logit Output, RHP Versus LHB, 405,861 Observations

Variable	Coefficient	Std. Error	z-Statistic	p-value
$\ln(L_o)$	−1.010535	0.016668	−0.632091	0.5273
$\ln(B_o)$	1.021111	0.011191	1.886360	0.0592
$\ln(P_o)$	0.989424	0.012117	−0.872868	0.3827

TABLE 6
Binary Logit Output, RHP Versus RHB, 441,000 Observations

Variable	Coefficient	Std. Error	z-Statistic	p-value
$\ln(L_o)$	−1.004607	0.015672	−0.294052	0.7687
$\ln(B_o)$	1.012817	0.010559	1.213828	0.2248
$\ln(P_o)$	0.991790	0.011288	−0.727387	0.4670

for any group of explanatory variables. For our purposes, an observation consists of the binary outcome of the plate appearance (strikeout or not strikeout), the batter and pitcher strikeout and ground ball rates for the year and platoon configuration, and the league averages (L and L_G) for the year and platoon configuration. The rates and averages are computed using the process described in Section 3.1.2.

4 Generalized Log5 Model

Starting from the logit model of Equation (10), we can consider generalized log5 models of the form

$$\ln(E_o) = c_0 \ln(L_o) + c_1 \ln(B_o) + c_2 \ln(P_o), \tag{13}$$

where the constant c has been written $c_0 \ln(L_o)$ to facilitate comparison with the log5 model and to make explicit that this term varies from year to year as the league average L varies. The coefficients c_0, c_1, and c_2 can be estimated for each platoon configuration using logistic regression as described in Section 3.2.

In Section 2.1 we observed that the log5 model has the desirable property that $E^* = L$ if $B = P = L$. The model of equation (13) will also have this property if $c_0 + c_1 + c_2 = 1$. Thus, we use logistic regression with this constraint to estimate the value of the coefficients for each platoon configuration. Tables 3, 4, 5, and 6 present the output of the

TABLE 7
Mean and Maximum of Absolute Difference between E_3 and E^*

Pit_Hand	Bat_Hand	Observations	Mean($\lvert D_3\rvert$)	Max($\lvert D_3\rvert$)
Left	Left	24,419	0.001816	0.008277
Left	Right	124,992	0.001533	0.014289
Right	Left	405,861	0.000950	0.006218
Right	Right	441,000	0.000633	0.005897

constrained logistic regression. Each table includes the value of the coefficients and their standard error along with the corresponding z-statistics and p-values computed using the log5 model coefficients $(c_0 = -1.0, c_1 = 1.0, c_2 = 1.0)$ as the null hypothesis. The p-values are greater than 0.05 for each variable for each platoon configuration which generally supports the use of the standard log5 model for this application. We observe, however, that for the left-handed pitcher versus right-handed batter configuration (Table 4), two of the variables have p-values that are below 0.08 which suggests that an improved model in the form of (13) may be possible for this case. We will examine this platoon configuration in more detail later in this section.

Using the c_0, c_1, and c_2 coefficients, the estimated strikeout probability E_3 for the three-variable model of (13) for a plate appearance within a platoon configuration is given by

$$E_3 = P(K = 1|L, B, P) = \frac{1}{1 + e^{-S}}, \tag{14}$$

where

$$S = c_0 \ln(L_o) + c_1 \ln(B_o) + c_2 \ln(P_o) \tag{15}$$

and where L, B, and P are the strikeout rates for the league, batter, and pitcher that have been estimated for the year and platoon configuration as described in Section 3.1.2.

We can further quantify the difference between this generalized model and the log5 model by considering the deviation $D_3 = E_3 - E^*$ for each plate appearance that was used by the constrained logistic regression to create the model. Table 7 presents the average and maximum value of the absolute difference $|D_3|$ for all of the plate appearances that were considered for each platoon configuration.

For each platoon configuration except LHP versus RHB, the maximum absolute difference between E_3 and E^* is less than 1 percent. The difference surface $D_3(B, P)$ for the LHP versus RHB configuration for 2013 is plotted in Fig. 4 and we see that the absolute difference D_3 is the largest when both strikeout rates B and P are large.

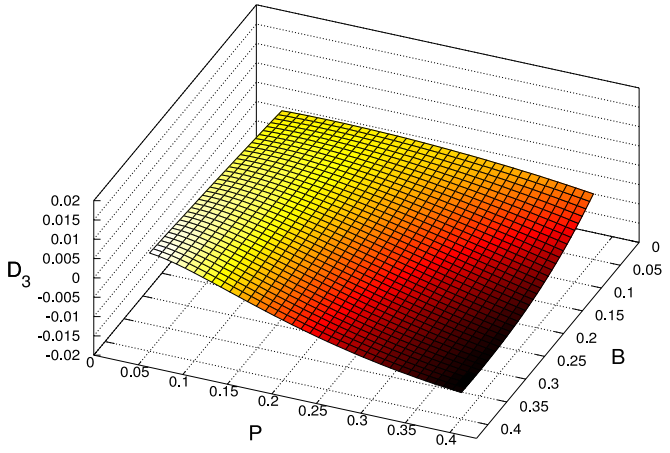


Fig. 4. $E_3 - E^*$ surface for LHP versus RHB, 2013.

TABLE 8
Descriptors for 2010 Matchup between Left-Handed Pitcher Billy Wagner and Right-Handed Batter Mark Reynolds

Player_ID	Role	Year	Configuration	SO Rate	GB Rate
wagnb001	Pitcher	2010	LHP versus RHB	0.396985	0.304762
reynm001	Batter	2010	LHP versus RHB	0.345679	0.358974

The largest absolute difference in our dataset occurs for a 2010 matchup between left-handed pitcher Billy Wagner and right-handed batter Mark Reynolds with the player descriptors shown in Table 8. This matchup gives a log5 E^* of 0.634525 and a generalized log5 E_3 of 0.620236 for the difference of $D_3 = -0.014289$. We note that both batter and pitcher strikeout rates for this case are much larger than the league average of $L = 0.167$ for this year and platoon configuration.

5 Incorporating Ground Ball Descriptors

5.1 Four-Variable Model

As explained in Section 3.1.1, batter and pitcher ground ball rates can be estimated with a high reliability within a platoon configuration using the data from a single season. Thus, we can ask whether the use of ground ball rates can improve the ability to model the probability of a strikeout for a matchup. For a given platoon configuration and year, let B_G, P_G, and L_G be the batter, pitcher, and league ground ball rates as defined in Section 3.1.2. After using logistic regression to consider a large set of candidate models that included log odds ratio and linear terms with cross terms in strikeout and ground ball rate, we arrived at the four-variable model

$$\ln(E_o) = c_0 \ln(L_o) + c_1 \ln(B_o) + c_2 \ln(P_o) + c_3 \widehat{B}_G \widehat{P}_G, \tag{16}$$

where $\widehat{B}_G$ and $\widehat{P}_G$ are the centered ground ball rates

$$\widehat{B}_G = B_G - L_G, \qquad \widehat{P}_G = P_G - L_G. \tag{17}$$

The model in (16) has the same form as the generalized log5 model of (13) with the added ground ball rate cross term $c_3 \widehat{B}_G \widehat{P}_G$. The constraint $E = L$ if $B = P = L$ is not appropriate for this model since the expected strikeout probability for a matchup between a pitcher and batter with league-average strikeout rates is not necessarily L since the ground

TABLE 9
Binary Logit Output, RHP Versus LHB, 405,861 Observations

Variable	Coefficient	Std. Error	z-Statistic	p-value
$\ln(L_o)$	−1.012410	0.016738	−0.741413	0.4584
$\ln(B_o)$	1.019923	0.011180	1.782024	0.0747
$\ln(P_o)$	0.988184	0.012098	−0.976692	0.3287
$\widehat{B}_G \widehat{P}_G$	2.288284	0.852856	2.683085	0.0073

TABLE 10
Binary Logit Output, RHP Versus RHB, 441,000 Observations

Variable	Coefficient	Std. Error	z-Statistic	p-value
$\ln(L_o)$	−1.008536	0.015784	−0.540709	0.5887
$\ln(B_o)$	1.013005	0.010539	1.233986	0.2172
$\ln(P_o)$	0.990718	0.011272	−0.823499	0.4102
$\widehat{B}_G\widehat{P}_G$	2.028065	0.698214	2.904648	0.0037

ball rates B_G and P_G can play a role in adjusting the expected strikeout probability.

Tables 9 and 10 present the results of the logistic regression for the RHP versus LHB and RHP versus RHB cases with the log5 model coefficients $(c_0 = -1.0, c_1 = 1.0, c_2 = 1.0)$ used to define the null hypothesis. Each of the first three variables $(\ln(L_o), \ln(B_o), \ln(P_o))$ has a p-value that is greater than 0.05 for each platoon configuration which supports the use of the log5 model coefficients for these variables. In addition, the unconstrained sum $c_0 + c_1 + c_2$ is nearly one for each case. We note that if the coefficients satisfy $c_0 + c_1 + c_2 = 1$ then the expected strikeout probability for a matchup between a batter and pitcher with league average strikeout rates and league average ground ball rates will be the league average strikeout rate which matches the log5 prediction for this input. The ground ball rate cross term is highly significant with a p-value of 0.0073 for the RHP versus LHB configuration and 0.0037 for the RHP versus RHB configuration. The p-value for the ground ball rate cross term was not significant for the two platoon configurations involving left-handed pitchers. We believe that this is due to having many fewer observations for these cases.

The estimated strikeout probability E_4 for this four-variable model for a plate appearance within a platoon configuration is given by

$$E_4 = P(K = 1|L, B, P, L_G, B_G, P_G) = \frac{1}{1 + e^{-S}}, \tag{18}$$

where

$$S = c_0 \ln(L_o) + c_1 \ln(B_o) + c_2 \ln(P_o) + c_3 \widehat{B}_G \widehat{P}_G \tag{19}$$

and where L, B, P, L_G, B_G, and P_G are the rates estimated for the year and platoon configuration as detailed in Section 3.1.2.

We can examine the difference between E_4 and the log5 model by computing $D_4 = E_4 - E^*$ for every plate appearance observation that was used to build the model. Table 11 presents the average and maximum value of the absolute difference $|D_4|$ over the plate appearance observations for the RHP versus LHB and RHP versus RHB platoon configurations for which all four variables were significant. We see that these differences are larger than for the generalized

TABLE 11
Mean and Maximum of Absolute Difference between E_4 and E^*

Pit_Hand	Bat_Hand	Observations	Mean($\lvert D_4\rvert$)	Max($\lvert D_4\rvert$)
Right	Left	405,861	0.001625	0.033392
Right	Right	441,000	0.001566	0.028794

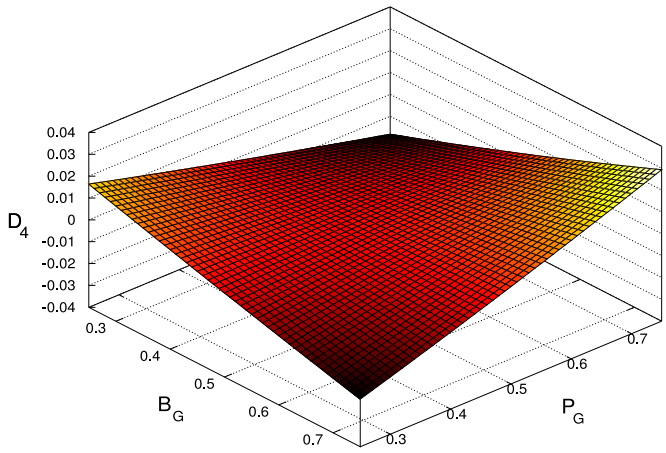


Fig. 5. $E_4 - E^*$ surface for RHP versus RHB, 2013, $(B = P = L)$.

log5 model considered in Section 4 and that the largest differences are now a few percent in predicted strikeout rate.

Figs. 5 and 6 allow us to examine the impact of the ground ball rate cross term on the predicted strikeout rate for the four-variable model as compared to the log5 model. The surface in Fig. 5 plots the difference $D_4 = E_4 - E^*$ as a function of the batter and pitcher ground ball rates with the batter and pitcher strikeout rates set to the league average $(B = P = L)$ for the RHP versus RHB platoon configuration for year 2013. The shape of the surface as defined by the model will be similar for other platoon configurations and years.

Fig. 6 plots the one-dimensional curves that result from intersecting the surface in Fig. 5 with the orthogonal planes $B_G = P_G$ and $B_G = 1 - P_G$. Since the c_0, c_1, and c_2 coefficients are similar for E_4 and E^*, equation (16) predicts and the plots illustrate that D_4 will be near zero when B_G and P_G are near the league average $(L_G = 0.463)$ for this platoon configuration and year. As we move away from the central area of the surface, however, E_4 becomes larger than E^* if we move along the direction $B_G = P_G$, but E_4 becomes smaller than E^* if we move along the orthogonal direction $B_G = 1 - P_G$. In other words, for cases where both B_G and P_G are distant from L_G, we will see more strikeouts if the batter and pitcher have similar ground ball rates and we will see fewer strikeouts if the batter and pitcher have significantly different ground ball rates.

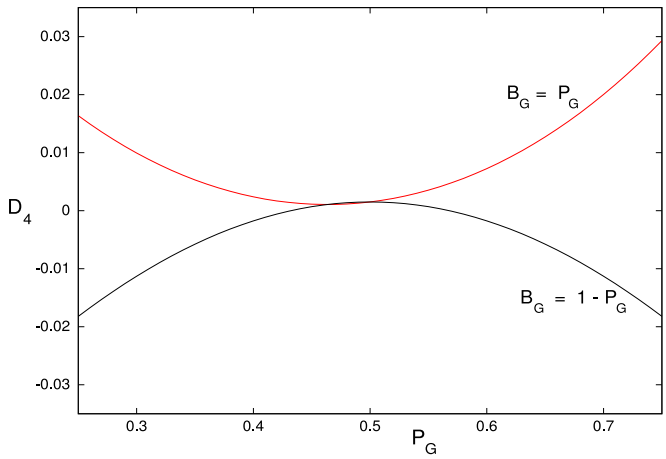


Fig. 6. One-dimensional slices of $E_4 - E^*$ surface (RHP versus RHB, 2013).

5.2 Physical Justification

The effect of the ground ball rate cross term as depicted in Figs. 5 and 6 is consistent with physical intuition. In baseball parlance, pitchers with high values of P_G are referred to as ground ball pitchers and pitchers with small values of P_G are referred to as fly ball pitchers. Similarly, batters with high values of B_G are known as ground ball hitters and batters with small values of B_G are called fly ball hitters. Ground ball pitchers tend to miss under bats and ground ball hitters tend to swing over pitches. The errors accumulate when a ground ball pitcher faces a ground ball hitter which leads to more swings and misses and a higher probability of a strikeout. Similarly, fly ball pitchers tend to miss over bats and fly ball hitters tend to miss under pitches. Therefore, the errors also accumulate when a fly ball pitcher faces a fly ball hitter which leads to more swings and misses and more strikeouts. On the other hand, if we consider a confrontation between a ground ball pitcher and a fly ball hitter, the ground ball pitcher tends to miss under bats while the fly ball hitter tends to swing under pitches. For this case, the errors cancel each other which leads to fewer strikeouts. Similarly, matchups between a fly ball pitcher and a ground ball hitter will also lead to fewer strikeouts. Tango et al. [19] showed that these physical predictions were consistent with the overall results in matchups without specifically considering the case of strikeouts.

5.3 Examples

In this section, we present matchup examples from 2003-2013 that exhibit significant differences between the log5 and four-variable model predictions. Example 1 considers right-handed pitcher Samuel Deduno versus left-handed batter Jonathan Villar in 2013. Both pitcher and batter have a ground ball rate for this configuration that is significantly above the league average ($L_G = 0.435$) and the shape of the surface in Fig. 5 predicts that E_4 will be larger than E^* for this case. The output of the models gives $E_4 = 0.355064$ and $E^* = 0.321672$ for a difference of $D_4 = 0.033392$.

EXAMPLE 1

Samuel Deduno versus Jonathan Villar, $D_4 = 0.033392$

Player_ID	Role	Year	Configuration	SO Rate	GB Rate
dedus001	Pitcher	2013	RHP versus LHB	0.181467	0.641304
villj001	Batter	2013	RHP versus LHB	0.335404	0.703297

Example 2 provides a similar case from 2006 for a RHP versus RHB matchup where both pitcher and batter have a ground ball rate that is significantly above the league average ($L_G = 0.439$). Again, E_4 is a few percent larger than E^*.

EXAMPLE 2

Brandon Webb versus Yorvit Torrealba, $D_4 = 0.028794$

Player_ID	Role	Year	Configuration	SO Rate	GB Rate
webbb001	Pitcher	2006	RHP versus RHB	0.212919	0.701587
torry001	Batter	2006	RHP versus RHB	0.230337	0.680000

Examples 3 and 4 present matchups from the RHP versus LHB and RHP versus RHB configurations where both the pitcher and batter ground ball rates are well below league average. As predicted by Fig. 5, E_4 is larger than E^* for these cases.

EXAMPLE 3

Jon Papelbon versus Carlos Pena, $D_4 = 0.024027$

Player_ID	Role	Year	Configuration	SO Rate	GB Rate
papej001	Pitcher	2009	RHP versus LHB	0.293333	0.230769
penac001	Batter	2009	RHP versus LHB	0.297222	0.258883

EXAMPLE 4

Rafael Betancourt versus Rod Barajas, $D_4 = 0.024680$

Player_ID	Role	Year	Configuration	SO Rate	GB Rate
betar001	Pitcher	2010	RHP versus RHB	0.384106	0.219780
barar001	Batter	2010	RHP versus RHB	0.144578	0.189054

Examples 5 and 6 present matchups from the RHP versus LHB and RHP versus RHB configurations where the pitcher and batter ground ball rates both differ significantly from the league average with one rate above the average and the other rate below. As shown in Fig. 5, this leads to the E^* prediction being larger than E_4.

EXAMPLE 5

Ernesto Frieri versus Jonathan Villar, $D_4 = -0.029783$

Player_ID	Role	Year	Configuration	SO Rate	GB Rate
friee001	Pitcher	2013	RHP versus LHB	0.380368	0.222222
villj001	Batter	2013	RHP versus LHB	0.335404	0.703297

EXAMPLE 6

Chad Bradford versus Frank Thomas, $D_4 = -0.019243$

Player_ID	Role	Year	Configuration	SO Rate	GB Rate
bradc001	Pitcher	2003	RHP versus RHB	0.255000	0.718519
thomf001	Batter	2003	RHP versus RHB	0.195789	0.262987

6 Allocating the Variance in Expected Strikeout Rate

Using the models in this paper, we can estimate the fraction of the variance in the expected strikeout rate that is due to the batter and the pitcher. The four-variable model for expected strikeout probability that was presented in Section 5.1 can be written

$$E(\mathbf{v}) = \frac{1}{1 + e^{-S(\mathbf{v})}}, \tag{20}$$

where $\mathbf{v} = (v_1, v_2, v_3, v_4)$ is the vector with elements

$$v_1 = \ln(B_o), \quad v_2 = \ln(P_o), \quad v_3 = \widehat{B}_G, \quad v_4 = \widehat{P}_G \tag{21}$$

and $S(\mathbf{v})$ is defined by

$$S(\mathbf{v}) = c_0 \ln(L_o) + c_1 v_1 + c_2 v_2 + c_3 v_3 v_4. \tag{22}$$

We note that the log5 and generalized log5 models are a special case of Equations (20)-(22).

Using a first-order Taylor series, the variance of E can be approximated by [1]

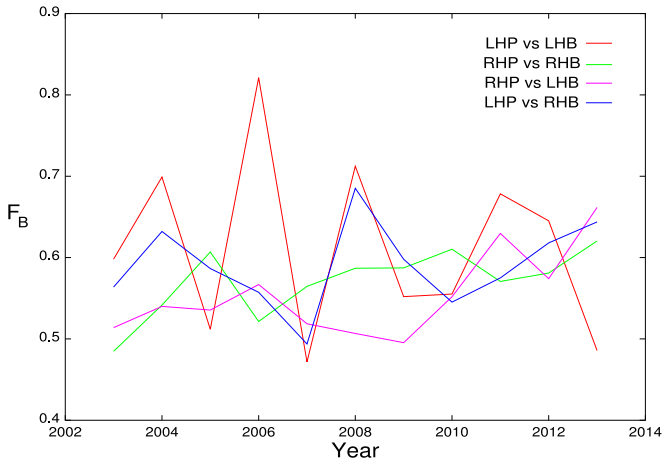


Fig. 7. Fraction of the variance in strikeout rate due to the batter.

$$\widehat{\sigma}_E^2 = \sum_{i=1}^{4} \left(\frac{\partial E}{\partial v_i} \right)^2 \sigma_i^2, \tag{23}$$

where σ_i^2 denotes the variance of v_i and the derivatives are evaluated at the mean $\overline{\mathbf{v}}$ of $\mathbf{v}$. Since the derivative $dE(S)/dS$ of the logistic function is $D(S) = E(S)(1 - E(S))$ the derivatives in (23) are given by the chain rule as

$$\begin{aligned} \frac{\partial E}{\partial v_1} &= D(S)c_1, \quad & \frac{\partial E}{\partial v_2} &= D(S)c_2, \\ \frac{\partial E}{\partial v_3} &= D(S)c_3 v_4, \quad & \frac{\partial E}{\partial v_4} &= D(S)c_3 v_3. \end{aligned} \tag{24}$$

On evaluation at $\mathbf{v} = \overline{\mathbf{v}}$, the last two terms in the sum of (23) are zero since v_3 and v_4 are zero-mean. Thus, the fractions of $\widehat{\sigma}_E^2$ that are due to the batter and pitcher respectively are given by

$$F_B = c_1^2\sigma_1^2 / \left(c_1^2\sigma_1^2 + c_2^2\sigma_2^2 \right) \tag{25}$$

$$F_P = 1 - F_B. \tag{26}$$

We computed the sample variances of v_1 and v_2 to approximate σ_1^2 and σ_2^2 for each platoon configuration and year in our study using all of the plate appearances for which reliable rates could be estimated according to the criteria described in Section 3.2. Fig. 7 plots the estimated F_B fraction for each platoon configuration and year using Equation (25) with the generalized log5 model coefficients. We see that batters account for most of the variance in expected strikeout rate and the average fractions over year for the four platoon configurations are given in Table 12. As expected, the configurations that involve left-handed pitchers contain more year-to-year variation since they are based on fewer plate appearances.

TABLE 12
Average Fraction of Variance in E_3 Accounted for by Batter and Pitcher

Pit_Hand	Bat_Hand	Mean(F_B)	Mean(F_P)
Left	Left	0.611876	0.388124
Left	Right	0.590771	0.409229
Right	Left	0.553999	0.446001
Right	Right	0.570502	0.429498

7 Conclusion

We have answered the questions posed in Section 3.1 by using 11 years of major league baseball play-by-play data to investigate models for the probability of a strikeout for a batter/pitcher matchup. Players are modeled using strikeout and ground ball rate descriptors that can be estimated reliably using data for a single platoon configuration for a single season. We used a constrained three-term logit model to show that the log5 formula provides an accurate model for strikeout probability and that small changes to the log5 coefficients might be used to improve the accuracy of the model for the LHP versus RHB platoon configuration. We also showed that a batter/pitcher ground ball rate interaction variable is highly significant when added to the three-term logit model. This interaction variable has a strong physical justification and adjusts the predicted strikeout probability based on the relative ground ball versus fly ball tendencies of the batter and pitcher. The models were used to show that batters are responsible for most of the variance in the predicted strikeout probability. The method employed to extend the log5 model to include additional variables can easily be adapted for other application areas. This paper has focused on the development and evaluation of low-dimensional models for strikeout probability and a natural next step is to assess the utility of these models for prediction.

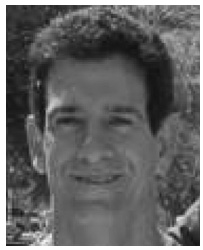
Acknowledgments

The data used in this study was obtained from www.retrosheet.org. The author would like to thank Tom Tango for his help in building a retrosheet database.

References

[1] A. H.-S. Ang and W. H. Tang, *Probability Concepts in Engineering Planning and Design*. New York, NY, USA: Wiley, 1984.
[2] R. Carleton. (Aug. 7, 2009). If you're happy and you know it, get on base [Online]. Available: www.hardballtimes.com/tht-live/if-youre-happy-and-you-know-it-get-on-base
[3] R. Carleton. (Jul. 16, 2012). It's a small sample size after all [Online]. Available: www.baseball.prospectus.com/article.php?articleid=17659
[4] R. Carleton. (May 9, 2013). Should I worry about my favorite pitcher? [Online]. Available: www.baseballprospectus.com/article.php?articleid=20516
[5] D. Fox. (Nov. 23, 2005). A short digression into log5 [Online]. Available: www.hardballtimes.com/a-short-digression-into-log5
[6] D. Fox. (Nov. 10, 2005). Tony LaRussa and the search for significance [Online]. Available: www.hardballtimes.com/tony-larussa-and-the-search-for-significance
[7] C. Hammond, W. Johnson, and S. Miller, "The James function," *Math. Mag.*, vol. 88, no. 1, pp. 54–71, Feb. 2015.
[8] B. James, *The Bill James Baseball Abstract*. New York, NY, USA: Ballantine Books, 1983.
[9] P. Jensen. (Jun. 30, 2009). Using HITf/x to measure skill [Online]. Available: www.hardballtimes.com/using-hitf-x-to-measure-skill
[10] T. Kline, *Psychological Testing: A Practical Approach to Design and Evaluation*. Thousand Oaks, CA, USA: Sage, 2005.
[11] A. Koo. (Dec. 18, 2013). More moneyball: Oakland's other platoon advantage [Online]. Available: www.baseballprospectus.com/art-icle.php?articleid=22435
[12] G. F. Kuder and M. W. Richardson, "The theory of the estimation of test reliability," *Psychometrika*, vol. 2, no. 3, pp. 151–160, 1937.

[13] D. Levitt. (Nov. 4, 1999). The batter/pitcher match up [Online]. Available: baseballthinkfactory.org/btf/scholars/levitt/articles/batter-pitcher-matchup.htm
[14] A. Miller, L. Bornn, R. Adams, and K. Goldsberry, "Factorized point process intensities: A spatial analysis of professional basketball," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 235–243.
[15] S. Staude. (Jun. 12, 2013). Better match-up data: Forecasting strikeout rate [Online]. Available: www.fangraphs.com/blogs/bettermatch-up-data-forecasting-strikeout-rate
[16] S. Staude. (Jun. 14, 2013). Batter-pitcher matchups part 2: Expected matchup K rate[Online]. Available: www.fangraphs.com/blogs/batter-pitcher-matchups-part-2-expected-matchup-k
[17] S. Staude, "Revisiting The Book's 'Mano a Mano' chapter," in *The Hardball Times Baseball Annual 2014* J. Distelheim, G. Simons, and P. Swydan, Eds. Fangraphs and The Hardball Times, USA, Lexington, KY, pp. 281–293, Nov. 2013.
[18] H. Stern and A. Sugano, "Inference about batter-pitcher matchups in baseball from small samples," in, *Statistical Thinking in Sports*, J. Albert and R. Koning, Eds. London, U.K.: Chapman & Hall/CRC, 2007, pp. 153–165.
[19] T. Tango, M. Lichtman, and A. Dolphin, *The Book: Playing the Percentages in Baseball*. Dulles, VA, USA: Potomac Books, 2007.
[20] Y. Yue, P. Lucey, P. Carr, A. Bialkowski, and I. Matthews, "Learning fine-grained spatial models for dynamic sports play prediction," in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 670–679.



Glenn Healey received the BSE degree in computer engineering from the University of Michigan and the MS degree in computer science, the MS degree in mathematics, and the PhD degree in computer science from Stanford University. He is currently a professor of electrical engineering and computer science at the University of California, Irvine. Before joining UC Irvine, he was at IBM Research. He was on the editorial boards of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, and the *Journal of the Optical Society of America*. He has received several awards for outstanding teaching and research. He is a fellow of the IEEE and the SPIE.