2022

Hot Boy Summer? Analyzing Managerial Reactions to Season-long Fluctuating Player Performance In Major League Baseball

John Rodgers Hood
Bowdoin College

Hot Boy Summer? Analyzing Managerial Reactions to Season-long Fluctuating Player Performance In Major League Baseball

An Honors Paper for the Department of Economics

By John Rodgers Hood

Bowdoin College, 2022

Abstract

This paper suggests numerical weights that a Major League Baseball (MLB) manager may use when comparing player performance across multiple past performance periods to predict future performance. By the end of the MLB regular season, current season performance becomes more predictive than prior season performance for pitchers but not hitters. After estimating weights for different past time periods of performance, this paper compares the weights with how managers value performance in high-stakes situations across these same time periods. I find that MLB managers overreact to recent performance by both hitters and pitchers in postseason settings.

Acknowledgements

I thank my parents for constantly telling me to be relentless in pursuing my passions. I am very grateful to receive the support they have given me, both morally and financially in my academic and athletic endeavors. The emotional, moral, and financial support I received from my closest family and friends throughout my life has been paramount to my work. I cannot emphasize the importance of conversation in my personal and academic development. My work habits have been shaped almost entirely by my peers. Throughout my time at Bowdoin, deep conversations with friends have yielded numerous breakthroughs in times of struggle.

I would especially like to thank Professor Daniel Stone, my Honors advisor, for his guidance and mentorship over the course of my time in college. His advising helped me figure out my economic interests. Over the 2021-2022 academic year, the consistent and clear guidance and advice he gave me yielded many of the ideas and analyses explored in this paper. His flexibility, intelligent thought, and creativity has greatly contributed to my personal and academic growth as an aspiring researcher.

I would also like to thank Professor Jessica LaVoice and Professor Matthew Botsch. As my readers, they each provided me with critical and insightful feedback that improved my work substantially. As my teacher, Prof. LaVoice developed my understanding for data analysis and econometrics in invaluable ways. Prof. Botsch's careful, nuanced considerations and questions challenged me to levels I had not previously experienced.

I claim responsibility for any and all errors in this Honors thesis.

Contents

1 Introduction

How do experienced actors make decisions when gifted with copious information? How do these actors make decisions in scenarios with significant consequences for others and themselves? As an actor gathers more information, does she update her beliefs in a Bayesian manner? How optimal are her decisions? Is decision optimality quantifiable?

To better understand how actors make high-stakes decisions using past information, I turn to the controlled setting of Major League Baseball (MLB). From April through August, each of the 30 teams in the MLB contains 26 players on its active roster and one manager. Beginning in September, the active roster size expands to 28. The manager is primarily responsible for deciding which players on the active roster play each game. There are nine positions and a manager must choose one player per position at any given time. Of these nine players, one takes the role of pitcher and the others play in the field. Fielders are typically evaluated primarily on their hitting performance, while pitchers are primarily evaluated on their pitching performance (despite hitting occasionally). In the case of American League (AL) games, the manager has the option to start a player as a team's designated hitter (DH), for a total of ten players playing simultaneously. The DH does not play in the field; instead, he hits in the pitcher's place. To distinguish between these types of players, I characterize all players who are not pitchers as hitters.

The situation where 26-28 players compete for nine or ten available playing positions introduces a scarcity issue. When such scarcity is considered, the manager's decision concerning which players to play (i.e. to take on the role of one of the available positions) and which players to bench (i.e. to *not* take on a role of one of the available positions) may be framed as a constrained optimization problem where the manager chooses players to maximize the probability of winning of a championship [Leitch, 2022].

Often, a manager encounters the following situation. He has two players on his team who play the same position who are relatively similar in ability. Perhaps one player has played better in past seasons, but the other player has played better in the current season. The manager may only start one player at a time. Which player should the manager start

over the next few games to maximize his chances of victory? This scenario frequently arises in high-stakes situations including winner-take-all playoff games. Here, the manager has substantial information about each player and must efficiently use this information to make an optimal decision. How should the manager optimally weigh player performance across different time periods to make the best decision possible?

A manager must properly balance short-term winning with long-term player development to maximize his team's probability of winning a championship. Each team has its own time horizon on which it expects and plans to win its next championship. Some teams believe they have a relatively high probability of winning a championship in the current season, while others believe they do not have the current ability to win a championship and shift their focus to winning a championship multiple years into the future. Teams in the former situation are more likely to focus on winning in the short-term, while teams in the latter situation are more likely to prioritize winning in the long-term. Teams in these two situations act differently by prioritizing long-term and short-term development to differing degrees. Teams prioritize long-term success tend to emphasize developing young players who have the potential to contribute to success multiple years in the future. On the other hand, teams prioritizing short-term success more frequently play the players that give them the best chance to win in the present. The MLB postseason is the time of season where teams most clearly prioritize winning in the short-term. Once a team qualifies for the postseason, it has a higher probability of winning a championship relative to its chances at other points in the season[1]. After qualifying for the postseason, teams have fewer days and opponents between them and winning a championship. At this point in the season, other teams have been eliminated from championship contention and each game carries significantly more weight than in the regular season. To qualify for the playoffs, a team must win a certain proportion of their 162 regular season games. Losing a single game here carries less significance than losing a playoff game. To win a championship, a team must win multiple best-of-seven game series and so postseason games hold higher stakes than regular season games, where losing is

[1]Each season, 8-10 teams, out of 30, qualified for the MLB postseason from 2000-2019

less costly. Here, I assume that prioritizing short-term winnings in the postseason is rational manager behavior. While managers may give playoff experience to younger players at the cost of short-term success, such a situation rarely occurs. Therefore, when evaluating manager decision-making optimality with respect to subsequent player performance, I evaluate the decisions of managers made between postseason games. In doing so, I assume that managers aim to maximize individual player performance at each position to maximize overall team performance.

In this paper, I conduct two distinct empirical exercises. First, I aim to determine how much value a manager should place on player performance over various periods of time. Raab and Gula [2004] describe the hot hand belief as "the belief that the performance of an athlete temporarily improves following a string of successes". I am interested in determining the existence of a season-level hot hand in baseball. More specifically, I aim to answer the following questions. How predictive is current season performance of future performance when compared against all given information a manager knows about a player? Such additional information includes previous performance and experience, age, recent performance, and opponent ability. How well does prior career performance predict subsequent performance in a given season relative to current season performance for different types of players at various points in the season? How and why might these results vary by position? More rigorously, I evaluate the validity of the hypothesis that after some point in the regular season, current season performance is more predictive than prior career performance and I hypothesize that current season performance adds predictive value when isolated from all other information available to the manager, such as a player's recent performance, his prior career performance, age, and other factors discussed in section 3.

I find that prior career performance tends to be more predictive of subsequent performance among hitters, while current season performance is more indicative of subsequent performance for pitchers. Upon exploring the hot-hand, I find a slight short-term hot hand or "short-term predictability in performance" effect consistent with Green and Zwiebel's [2018] findings that following strings of successes, a player experiences a temporary boost in per-

formance.

The second empirical exercise pertains to manager decision-making optimality. I aim to evaluate how well managers understand the predictive value of these variables, particularly the predictive value of performance across multiple time periods. After estimating the predictive value of multiple periods of performance, I compare these values to the weights managers implicitly assign to each of these periods when managers make decisions. I aim to answer the questions: How optimally do managers make decisions regarding which players to give playing time? How identifiable are such decision-making inefficiencies? I examine the possibility that managers overvalue or undervalue recent performance and undervalue or overvalue prior career performance, evaluating each question with respect to hitters and pitchers. I expect an overreaction to recent performance to be most likely and find results consistent with my expectations.

1.1 Significance

The results of this study have direct significance to actors in baseball and the general decision-making population. First, the results may inform decision-making tendencies and beliefs of baseball managers, front office members, fans, and players. Managers may gain direction concerning how to decide which players to start and which players to bench. Within an MLB team's front office, general managers and scouts may properly use prior information to estimate player value. With accurate forecasts, front offices may better determine which players to compensate and how much they should be paid. Fans may use the results to identify market inefficiencies in fantasy baseball and gambling while opposing players may adjust in-game strategy to account for the effects. After considering an opponent's prior performance over multiple time periods, pitchers and hitters may adjust how conservative or aggressive they act when facing opposing players. For example, once recognizing an opposing hitter is hot, a pitcher must determine how to appropriately act. Green and Zwiebel [2018] determine that pitchers tend to overreact to the short-term hot hand by walking hot players more frequently than what Green and Zwiebel [2018] consider optimal. This study aims to

inform how pitchers react to longer periods of an opponent's sustained success (or failure).

Anecdotal evidence of a season-long hot hand may appear most clearly in the case study of Baseball Hall of Famer Cal Ripken Jr.'s 1991 season [Rosenfeld, 1995]. Over the 1988-1990 seasons, Ripken Jr. averaged roughly 22 home runs per season and an 0.257 batting average. For context, 34 players hit more home runs that Ripken Jr. in 1990 and the MLB league average batting average over this period was 0.258 [199, a], suggesting Ripken Jr. to be an average hitter. In 1991, Ripken Jr. recorded 34 home runs and a 0.323 batting average, good for 4th and 6th in the MLB, respectively [199, b]. Ripken Jr. performed well above average, going on to win the AL Most Valuable Player award, maintaining above-average play throughout the season's duration. His play during the beginning of the season was indicative of his play in the second half of the season, particularly when compared to his poor prior performance (relative to the 1991 season). However, such strong performance was short-lived: the rest of his career, Ripken Jr. hit no more than 24 home runs in a single season. In the 1992-1993 seasons, his batting average regressed to 0.254. Such perplexity in performance fluctuation motivates my research. Was Ripken Jr.'s success in 1991 simply due to luck, or was he truly 'hot' over the entire regular season?

While most academic scholars are not overly concerned with baseball outcomes, baseball's controlled, high-stakes setting provides interesting case studies concerning decision-making. Stakes are high in multiple dimensions: each decision that a manager makes disseminates throughout the public via media press releases. Team operate as business organizations, paying managers multi-million dollar salaries each year to make decisions. These actions imply that their decisions hold significant value. These decisions are actions that people give ample attention. The MLB is a growing multi-billion dollar business with hundreds of millions of fans [Ozanian and Teitelbaum, 2022].

Performance-based outcomes have significance outside of baseball. Successes and failures achieved by individuals in various performance-based contexts including educational settings and the labor market may affect individual confidence levels. Descamps et al. [2022] find that a string of successes is more likely to follow an initial success than an initial failure. They

find such strings to be "driven by an information revelation effect, whereby players update their beliefs about their relative strength after experiencing an initial success" [Descamps et al., 2022]. If a reset in statistical recording at the beginning of each MLB regular season partially determine confidence levels influencing player performance, I may better understand the role that confidence effects play in settings broader than the MLB. This sample consists of managers who are (by definition) experienced actors with high stakes. Similar actors appear in financial settings. Recently, economists have studied belief formation, overreactions, and underreactions to news in macroeconomic settings with respect to different time periods. For example, Wang [2021] finds that U.S. Treasury bond market participants overreact to new information when forecasting short-term interest rates. On the other hand, these participants underreact to the same information when forecasting long-term interest rates. Similarly, Bordalo et al. [2019] find that stock market participants and analysts overreact to new information about a company when forecasting that company's future stock prices. The results of this study are relevant to Bayesian updating, decision-making, and the optimality of belief formation in contexts similar to these that focus on compiling substantial information to make accurate forecasts.

1.2 Additional related work pertaining to the hot hand

The short-term hot hand has been analyzed extensively in literature. Gilovich et al. [1985] wrote the canonical hot hand paper by claiming that no hot hand exists in basketball after analyzing the free throw attempts of college basketball players. However, advances in technology, computing power, and statistical methods have enabled researchers to conduct more rigorous analysis. After controlling for various exogenous factors, Miller and Sanjurjo [2018] and Miller et al. [2014] disprove Gilovich's claim by discovering a hot hand from the same data collected in Gilovich's original study. Stone [2012] finds measurement error and disputes Gilovich's claim, noting that by taking outcome-based approaches instead of a priori probabilistic approaches, a hot hand effect may not be found even when a significant hot hand effect exists. In light of these findings, Benjamin [2019] defines the hot hand

bias more generally to be the belief in a hot hand even though outcomes are known to be independent and identically distributed. Offerman and Sonnemans [2004] find evidence that decision-makers exhibit hot hand bias, overreacting to the hot hand in sports betting and stock market settings. On the other hand, Stone and Arkes [2018] discover an underreaction to the hot hand by the NCAA men's basketball selection committee when choosing teams to include in the NCAA men's basketball tournament and team seedings. In baseball, managers may be susceptible to hot hand (recency) bias. Miller and Sanjurjo [2019] discover a recency bias apparent in the decision-making of scientific researchers. They find that researchers analyze results similar to individuals facing the famous Monty Hall problem, with decision-making changing in light of recent streaks [see also Ayton and Fischer, 2004]. Game managers may act in a similar manner.

However, a change in scenery may halt certain streaks. Dai [2018] finds a reset effect after abrupt changes in scenery both generally and within baseball. More specifically, she finds that when traded from the National League to the AL, or vice-versa, streaky players reverse trends. Players who perform above their baseline previous performance before being traded to another league start to perform below this baseline. Similarly, players who perform poorly prior to being traded begin playing significantly better after being traded between leagues. Dai found evidence of a reversal effect when a player switched leagues. When the player stayed in the same league, however, the reversal effect was not found and the player maintained his pre-trade performance. By entering a new league, a player may clear his mind of poor prior performance. Once a player moves between leagues, his recorded statistics are often started anew. A fresh statistics record may gift the player with a clean slate, motivating him to take advantage of a new opportunity. Consistent with Dai's [2018] findings, I predict the start of a new season to act as a reset for player confidence and performance. A confidence effect may be present; once a player begins to play well in the new season, seeing above average statistics in the record books may give the player the confidence to keep playing well throughout the rest of the season.

Thus far, the literature has focused on short-term streakiness in player performance.

Season-long streaks of above and below average play have not been analyzed. The previous aforementioned studies have yet to take advantage of the wealth of data recently made available by increasingly frequent developments in sabermetrics and baseball analytics. This paper aims to expand on the current literature by measuring streakiness in baseball on a longer time scale than Green and Zwiebel [2018] use and evaluating how MLB managers react to these measures of streakiness using advanced metrics to measure player performance. Many metrics yielded by these developments have been shown to be more predictive of subsequent performance than traditional statistics [Richards, 2019].

1.3 Organization of paper

The organization of this paper is as follows. In section 2, I describe the sample data I use in the study and the criteria for a player to be included in a sample. In section 3, I describe the modeling and methodology I use in my first empirical exercise, predicting subsequent player performance, while section 4 displays my results and corresponding analysis concerning estimations of prior performance value. Section 5 explains the methodology I use to evaluate manager decision-making optimality and the corresponding results arising from the methodology. Finally, section 6 summarizes my key findings and concluding remarks.

2 Data

I use game-level statistics provided by *fangraphs.com* from 1295 pitchers and 1302 hitters playing in the 2003-2021 regular seasons. Each player has his own set of game logs which includes information by game: the date of the game, the player's team, opponent, number of plate appearances (batters faced for pitchers) and relevant hitting and pitching statistics to measure player performance. To explore manager decisions in the postseason, I collect additional game-level data from the 2006-2021 postseasons from *Baseball-Reference.com*. This data includes information concerning starting lineups and postseason player statistics.

Following the existing literature, I use expected fielding independent pitching (xFIP) to

measure pitching performance and weighted runs created plus (wRC+) to measure batting performance. The formula for xFIP is given below:

$$\text{xFIP} = \frac{\frac{13 \cdot \text{Fly balls}}{\text{league average rate of HR per fly ball}} + 3(\text{BB} + \text{HBP}) - 2\text{K}}{\text{IP}} + \text{FIP constant}$$

Where 'FIP constant' scales xFIP to the magnitude of earned runs allowed (ERA), a classical measure of pitcher performance, and:

- BB: number of batters walked.
- HBP: hit-by-pitch, the number of batters a pitcher hits with the ball.
- K: number of strikeouts.
- IP: innings pitched.

wRC+ measures hitting performance. According to Slowinski [2010a], wRC+ controls "for [ballpark] effects and the current season run environment. wRC+ is scaled so that the league average is 100 and each point above or below 100 is equal to one percentage point better or worse than the league average". wRC+ has controls that make it a better measure of hitting ability than classical metrics such as batting average, runs batted in (RBI), on base plus slugging (OPS), and weighted on base average (wOBA). wRC+ has a complicated formula; I direct the reader to [Slowinski, 2010a] for a mathematical definition.

3 Predictive model

First I estimate the effect of current season performance, *current*, on a player's subsequent performance, *next*. I measure performance across any particular time period for hitters and pitchers as described in section 2 and use multiple linear regression to evaluate current season performance's predictiveness concerning future performance over multiple periods

throughout the MLB regular season. At each point in time, I estimate the following model:

$$next = \beta_0 + \beta_1 last3 + \beta_2 current + \beta_3 recent + \gamma controls + \mu \tag{1}$$

where:

- *next*: player performance over the next $m \in \{25, 50, 75, 100\}$ player observations.
- *last3*: player performance over his previous three seasons.
- *recent*: player performance over his most recent $r \in \{25, 50\}$ player observations.
- *current*: player performance in the current season, up to the *recent* period.
- *controls*: player age, home field effects, opponent average performance and season, team, and season-team fixed effects.
- μ: error term

I use *last3* as a proxy variable to represent a player's baseline ability. By including three seasons of player observations in the baseline measurement period, I hope to precisely measure a player's baseline ability before the start of the current regular season. The more that a player plays in past seasons, the more precisely a manager may quantify his expectations regarding future performance. I measure *last3* as a straightforward average of player performance across his past three seasons. In appendix A.4 I show that my results do not change when $last3 = \alpha_1 l_1 + \alpha_2 l_2 + \alpha_3 l_3$, where l_i is a player's performance in the ith most recent season and each α_i is estimated by splitting *last3* into three distinct performance periods when estimating eq. (1).

I control for age, opponent average performance, home field effects, team fixed effects, year fixed effects, and team-year fixed effects. Determining how to best control for age is nontrivial because player improvement is nonlinear as a function of age. Younger players improve with experience, while older players lose skill through aging. To control for age I create a variable, $dfpeak = |age - 27|$ which measures a player's proximity to his peak age

27 [Hakes and Turner, 2011] under the belief that the closer a player is to his peak age, the better he plays, all else equal. For hitters, I measure opponent average performance as the opposing team's average xFIP in that year's regular season. For pitchers, I estimate opponent average performance to be the opposing team's average wRC+ over that year's regular season. I control for a player's percentage of home games in the *current*, *recent*, *last3*, and *next* player observation periods to implicitly capture ballpark and home field advantage effects [Jamieson, 2010]. I also include team, year, and team-year fixed effects. For a player to qualify for a sample, he must record at least 324 observations (an average of 2 per game) in each of the previous three seasons and at least $n + m$ player observations in the current season. Here, n is the number of player observations in the current season thus far. Table 1 shows each measurement period's length with respect to m, n, and r.

variable	first player observation	last player observation
current	1	n - r
recent	n - r + 1	n
next	n + 1	n + m

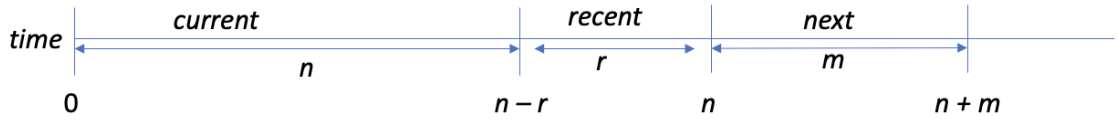

Table 1: Variables representing player performance across various periods of the MLB regular season, with accompanying visual. I list each period's first and last player observation.

For pitchers, I measure player observations by batter faced (BF). For batters, I define a player observation to be a plate appearance (PA). I estimate eq. (1) at multiple points in the season by varying $n \in \{100, 125, 150, ..., 500, 525, 550\}$, $r \in \{25, 50\}$ and $m \in \{25, 50, 75, 100\}$ with n, r, and m defined above. Each model I estimate can be represented by the parameterization (n, r, m), with separate estimations for hitters and pitchers.

To maintain simplicity, I hold $r = 25$ and $m = 100$ throughout this paper. I choose $r = 25$ because this is the period over which Green and Zwiebel [2018] found a short-term hot hand effect. I measure time by PAs and BFs instead of days to put players with varying levels of playing time on the same scale. For instance, one player may reach the 100 PA threshold in April while another may reach the same threshold in late June. If I measured

periods using days, the model would estimate equal effects of current performance on the subsequent performance of two players who have played vastly different amounts over the same number of days. By using PAs (BFs), I aim to capture the effect of current season performance $n - r$ PAs (BFs) into the season. By estimating eq. (1) at myriad thresholds, I aim to compare players who compete for playing time at the same position who have played different amounts in the season. This measurement strategy allows me to compare starters and substitutes, the latter of which may receive significantly less playing time than the former over a full season. By using PA and BF to measure time and estimating eq. (1) using samples with low thresholds of playing time, I aim to include substitute players in at least one of the samples so that I may predict their subsequent performance.

I also vary the length of *current* to analyze its estimated effect on *next* relative to *last3* at different points in the season. Longer periods include more player observations (by definition). Therefore, longer periods enable better estimations of player performance in that time period as smaller sample sizes of player observations tend to have larger standard deviations. Longer measurement periods help eliminate a proportion of this noise and give a more precise measure of player performance by increasing the amount of available information. For example, a player's performance over the first 500 PA of the current season should provide more information about the player's ability than his performance over the first 5 PA.[2]

Recall β_1 and β_2, the coefficients in eq. (1) corresponding to *last3* and *current*, respectively. I test the hypothesis $H_0 : \beta_1 = \beta_2$ against the alternative hypothesis $H_1 : \beta_1 \neq \beta_2$ for each combination (n, r, m) to evaluate the effect of current season performance on subsequent performance relative to prior performance at different points throughout the season. H_0 states that a player's current season performance and his performance over the last three seasons hold the same predictive power when estimating the his subsequent performance. Alternatively, H_1 finds one performance period more predictive than the other.[3]

[2]Information gained from plate appearances is not additive. 500 PA should not provide 100 times more information than 5 PA. I expect to gain an amount closer to $\sqrt{500/5} = 10$ times as much information.

[3]This paper focuses on the relationship between current and prior season performance rather than recent player performance. I refer the reader to Green and Zwiebel's work for information concerning β_3.

3.1 Sample selection

The raw data includes game-level statistics on players from the 2003 to 2021 MLB seasons. Hence, I define a model observation to be by player, by season, for one observation per player per season. I restrict a player to have at most one model observation per season. For example, when $n = 100$, I include a player who records 500 plate appearances in the given season only once. The sample of players I use in my analysis varies for each (n, r, m) parameterization defined in section 3. For a player-season combination to be included in a sample, a player must record at least 324 (average two player observations per game for 162 games, the length of the MLB regular season) PA or BF in each of his previous three seasons and enough PA or BF in the current season to calculate subsequent performance. For example, for a hitter's 2016 season to qualify for the $(525, 25, 100)$ sample, he must have recorded at least 324 PA in each of the 2013, 2014, and 2015 MLB regular seasons and at least $n + r + m = 525 + 25 + 100 = 650$ PA in 2016.

A direct consequence that follows from my sample selection methodology is that players without sufficient playing time do not qualify for the sample. As the current season playing time threshold increases, more players are excluded from the sample. Therefore, each sample's size decreases as the threshold increases. The sample $(75, 25, 25)$, with a threshold of 125 PA, includes 1857 player-season combinations while the sample $(525, 25, 100)$, with a threshold of 650 PA, includes 399[4] (after accounting for missing values). Therefore, I expect estimated coefficients to be less precise for smaller sample sizes. This expectation is consistent with my findings. I show that estimated standard errors increase as sample sizes decrease in section 4. To counter this increase, I estimate eq. (1) for various n and demonstrate that when I hold the sample constant, my results do not change.

[4]See section appendix A.5 for summary statistics.

4 Season-level hot hand results

4.1 Hitters

Before showing results, I clarify my choice of notation with an example by displaying hitter regression results for $(n, r, m) = (525, 25, 100)$. Since the model observations are hitters, a player observation is a PA. The *current* player observation period includes PAs 1 to 525, the *recent* PA period includes PAs 526-550, and the *next* PA period includes PAs 551-650. I show relevant estimated coefficients in Table 2. The estimated coefficients on *last3* and *current* imply that *last3* and *current* hold similar predictive value at the specified point in the season. These estimated coefficients appear in Figure 1: the right-most blue circle represents the estimated coefficient on *current* and the right-most red square represents the estimated coefficient on *last3*.

Table 2: Hitter regression results for $(n, r, m) = (525, 25, 100)$. The first column describes the independent variable. The second column represents the variable's estimated coefficient and corresponding heteroskedastic-robust standard error (in parentheses below). Additional regression statistics lie below the estimated coefficients.

	Dependent variable: next
last3	0.448^{***}
	(0.126)
current	0.427^{***}
	(0.109)
recent	-0.018
	(0.027)
Observations	399
Adjusted R^2	0.167
Note:	*p<0.1; **p<0.05; ***p<0.01

To be included in this regression, a player must have achieved at least 324 PA in each of the previous three seasons and at least 650 plate appearances in the current season. There are 399 player-season combinations that meet this threshold. The model explains 25 percent of the variance in *next* and the difference between estimated coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ on *last3*

and *current* is not statistically significant.

I plot the estimated coefficients on *last3* and *current*, $\hat{\beta}_1$ and $\hat{\beta}_2$, in Figure 1 from regressions performed throughout the season.

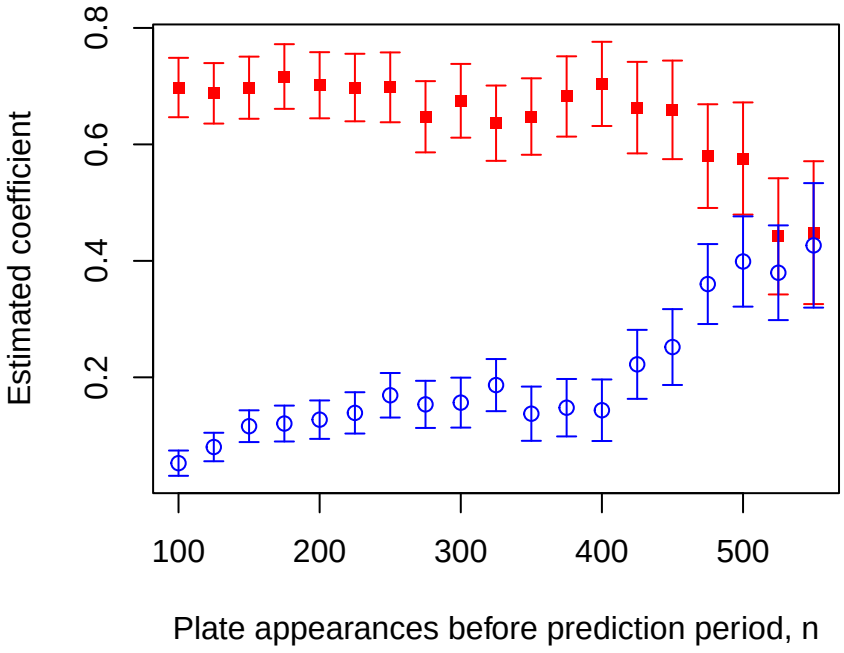

Figure 1: Hitters: estimated coefficients of the *current* (as blue circles) and *last3* (as red squares) variables in eq. (1) by plate appearance. Error bars represent heteroskedastic-robust standard errors.

The horizontal axis records n, the number of plate appearances recorded up to the point of prediction in the current season. The *next* period has length $m = 100$ PA with *recent* period length $r = 25$ PA.

Hence, Figure 1 shows the estimated coefficients as I vary n from 75 PA to 525 PA by increments of 25 with parameters $(n, r, m) = (n, 25, 100)$. Early in the season, prior career performance has a strong effect relative to current season performance. For small n, a hitter's current season performance will be noisy. Under the assumption that the outcomes of plate appearances are independent and identically distributed random variables, the weak law of large numbers implies that player performance may vary significantly in the short-term before

reflecting a player's true ability. As the length of the current season grows, a player's current season performance should become a more reliable measure of the player's ability. This increase in reliability should be reflected in predicting his subsequent performance (which I measure using *next*). In Figure 1, $\hat{\beta}_1 \in (0.6, 0.8)$ and $\hat{\beta}_2 \in (0, 0.2)$ from $1 \leq n \leq 425$. When $n > 425$, *current* gains predictive ability: $\hat{\beta}_2$ increases to nearly 0.4 by 475 plate appearances. Similarly, $\hat{\beta}_1$ decreases to roughly 0.4 at $t = 475$ PA.

In Figure 1, I allow the sample selection threshold to vary across different values of n to estimate eq. (1). I also aim to evaluate the effect of holding the set of player-season combinations constant while changing the number of plate appearances thus far in the current season. I hold the sample constant by requiring all player-season combinations to reach at least 600 PA for all $(n, 25, 100)$ samples and show the results in Figure 2.

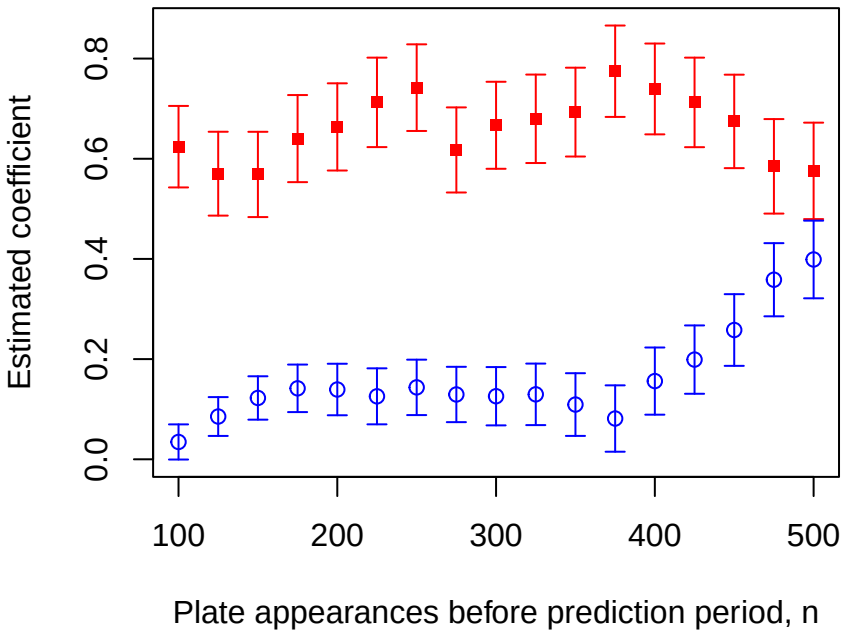

Figure 2: Hitters: estimated coefficients of the *current* (as blue circles) and *last3* (as red squares) variables in eq. (1) by plate appearance. Error bars represent heteroskedastic-robust standard errors. Each coefficient is estimated on the same fixed sample of player-season combinations with sufficient observation to qualify for the $n = 500$ sample.

Similar to results shown in Figure 1, Figure 2 shows estimated coefficients on *current* and *last3* holding constant until roughly 400 PA. When a hitter reaches the 400 PA mark, the predictiveness of *current* increases and the predictiveness of *last3* decreases. These trends are consistent with my original findings.

4.2 Pitchers

Keeping $r = 25$ and $m = 100$, I show pitcher regression results for models $(n, 25, 100)$ for $n \in \{75, 100, 125, ..., 500, 525\}$ in Figure 3.

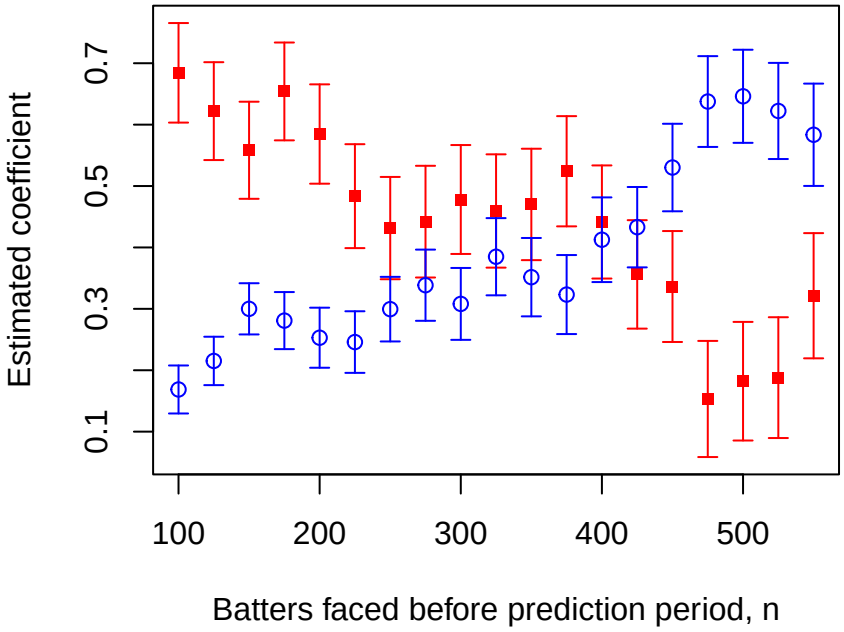

Figure 3: Pitchers: estimated coefficients of *current* (as blue circles) and *last3* (as red squares) in eq. (1) by batters faced. Error bars represent heteroskedastic-robust standard errors.

The horizontal axis records n, the number of batters faced up to the prediction point in the current season. The *next* period has length $m = 100$ BF with *recent* period length $r = 25$ BF. These are the same parameters I use to show hitting regression results. Once

again, I use $r = 25$ to control for the hot hand effect that Green and Zwiebel [2018] find players exhibit from their most recent 25 BF. Similar to the hitting regression results shown in fig. 1, prior career performance holds more weight than current season performance early in the season. $\hat{\beta}_2$ steadily increases from 0.174 at 100 BF to 0.658 at 475 BF. I also see a steady decrease in $\hat{\beta}_1$ over the same period (0.736 to 0.147). After $BF = 475$, I find a trend reversal. $\hat{\beta}_1$ decreases slightly while $\hat{\beta}_2$ increases slightly. Small sample sizes deriving from a small number of players meeting the qualification criteria late in the regular season may be an underlying factor causing this trend reversal. Therefore, I re-estimate eq. (1) at the same points $(n, 25, 100)$ for $n \in (100, 125, ..., 500)$ using only players who record at least 600 BF in the current season. Therefore, the sample remains constant while the number of BF varies. I show the results in Figure 4 and do not find any significant differences between the results shown in Figure 3 and Figure 4.

The estimated coefficient on *current* rises and the estimated coefficient on *last3* decreases gradually as n increases. Both figures demonstrate that *current* holds significantly more predictive value than *last3* once a player has faced 500 batters in the current season. In other words, previous season performance is significantly less predictive of subsequent performance than current season performance late in the season.

4.3 Controlling for future performance

In evaluating player performance for predictive purposes, it is optimal for general managers to differentiate between fundamental changes in baseline player ability and a long stretch of streaky play. A player may be having a particularly strong season for (at least) two reasons. On one hand, he may have improved his overall ability in which his statistics accurately reflect this improvement. For example, he may be stronger, faster, or improved his hand-eye coordination. On the other hand, he may be 'hot' in the current season. When predicting future performance, correctly distinguishing between these two causes gains importance if the hot player cools down in-season. I control for improvement by adding *future*, the player's performance in the subsequent season, as an explanatory variable to equation

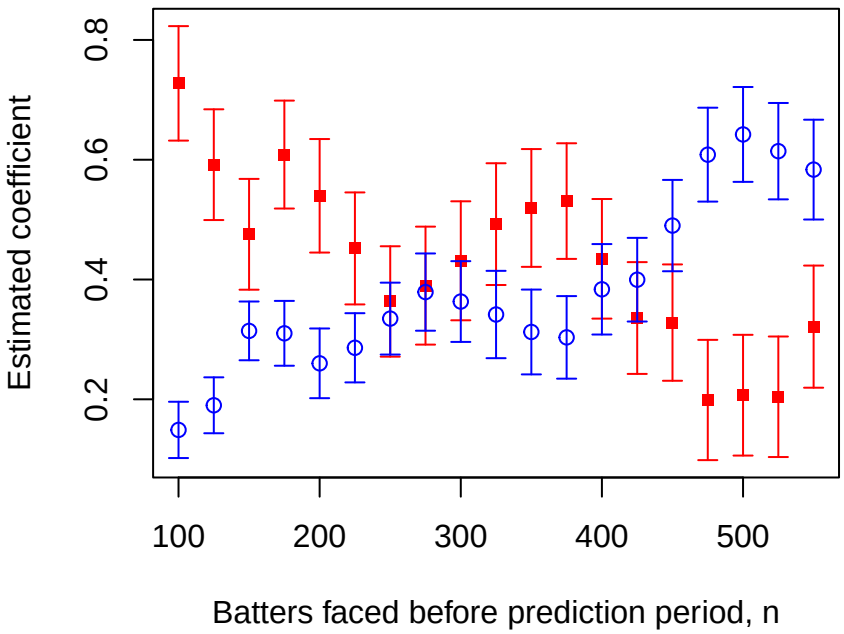

Figure 4: Pitchers: estimated coefficients of *current* (as blue circles) and *last3* (as red squares) in eq. (1) by batters faced. Error bars represent heteroskedastic-robust standard errors. Each coefficient is estimated on the same fixed sample of player-season combinations with sufficient observation to qualify for the $n = 500$ sample.

1. If a performance fluctuations are caused by fundamental changes in ability, the estimated coefficient $\hat{\beta}_4$ on future should be positive, as such a robust change in performance should be reflected over substantial periods of time. On the other hand, if a change in performance is caused by player streakiness that may reset at the season's duration [Dai, 2018], *future* performance should not add predictive information.

I present results with *future* as a control variable in Figure 5 and Figure 6. I compare Figure 5 and Figure 6 to Figure 1 and Figure 3 respectively and find no significant change in estimating β_1 and β_2 for both hitters and pitchers.

Qualitatively, Figure 5 demonstrates the same dynamic trends: for hitters, $\hat{\beta}_2$ increases and $\hat{\beta}_1$ decreases at the same points in the season as found previously. However, unlike the results shown in Figure 1, $\hat{\beta}_2$ does not surpass $\hat{\beta}_1$ at any point in the regular season.

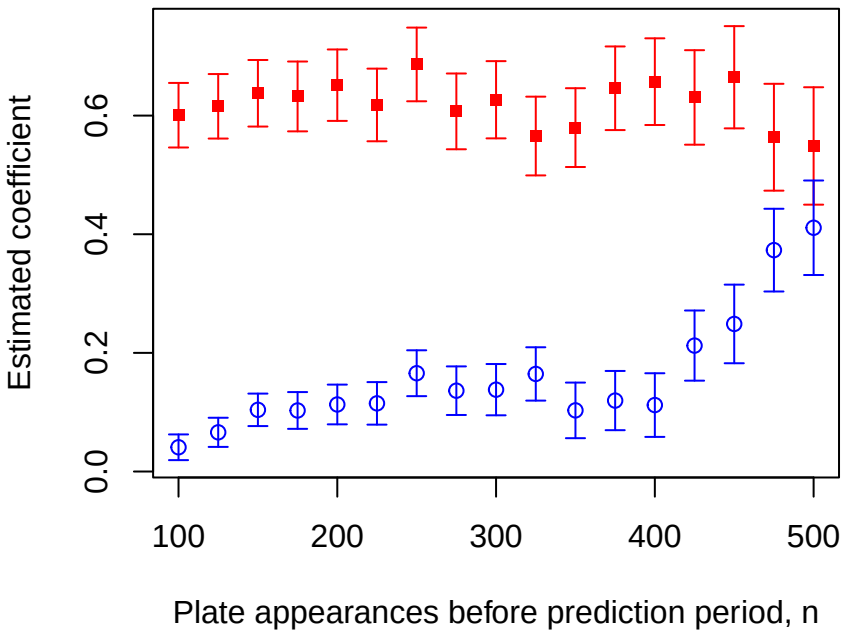

Figure 5: Hitters: estimated coefficients of *current* (as blue circles) and *last3* (as red squares) in eq. (1) by plate appearances with future performance included as an explanatory variable. Error bars represent heteroskedastic-robust standard errors.

Hence, when controlling for future season performance, *current* loses predictive value relative to *last3*. Such a decrease in predictive value may imply that future performance is more correlated with current performance than a player's performance over his last three seasons. These results are consistent with Dai's [2018] findings regarding resets and their effects on fluctuation in performance. The more resets that occur between performance periods, the more likely one period's performance is to substantially deviate from the other.

Figure 6 shows blue and red trend lines more difficult to identify than those shown in Figure 3. However, $\hat{\beta}_1$ and $\hat{\beta}_2$ maintain decreases and increases over the course of the season to values consistent with the estimated coefficients shown in Figure 3.

From the manager's point of view, determining the cause of improved play as either streakiness or intrinsic player improvement is unimportant in the short-term. Either way, increased player performance is desired. As mentioned in section 1, this distinction may be

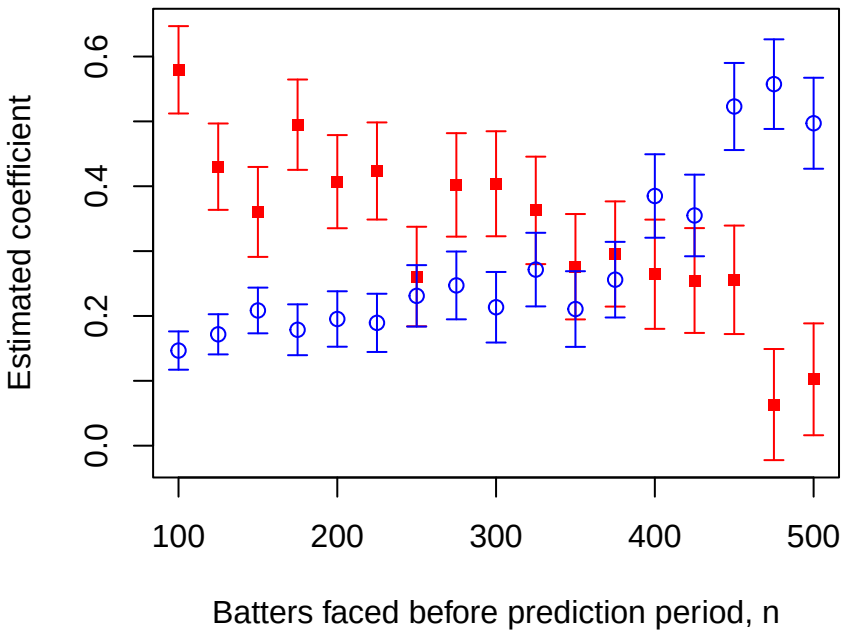

Figure 6: Pitchers: estimated coefficients of *current* (as blue circles) and *last3* (as red squares) in eq. (1) by batters faced with future performance included as an explanatory variable. Error bars represent heteroskedastic-robust standard errors.

useful for general managers who deciding which players to sign to high-paying contracts. Paying a hot player likely to cool down the same as an improved player will be costly if the two players performed similarly over the same time period. Distinguishing between these two types of above-average play may be crucial to team success.

4.4 Selection bias

My sample selection methodology presents a bias problem. Certain groups of players may not qualify for samples with high player observation thresholds. For example, managers may bench players before they receive enough PA or BF to qualify for a sample because these players play poorly early in the season. The sample may be biased towards an above-average population, misrepresenting the MLB player population. The bias presents itseldf

in each sample, particularly among hitters: the average wRC+ of each sample is higher than the league average of 100 and increases as the season continues and the sample shrinks. Therefore, I restrict my sample to 'above average' players as characterized by fangraphs.com. For hitters, a player may only be included in a sample if he reaches the threshold PA (as defined in section 3.1) and averages a wRC+ greater than 115 over his last three seasons. Similarly, pitchers must average an xFIP less than 3.5 across his last three seasons. I expect above average performance in a player's most recent three seasons recorded over a sufficient number of player observations to establish a player as a steady starter and show the results corresponding to the refined samples in Figure 7 and Figure 8.

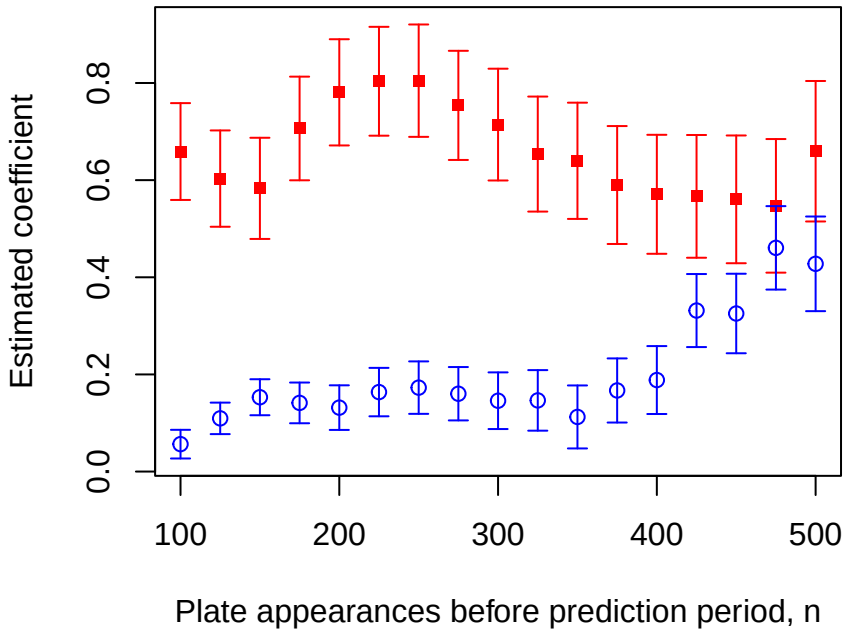

Figure 7: Restricted hitter sample: estimated coefficients of *current* (as blue circles) and *last3* (as red squares) in eq. (1). Error bars represent heteroskedastic-robust standard errors.

As compared to results shown in Figure 1, *last3* increases in predictive ability and *current* decreases in predictive ability among hitters. Figure 7 shows that $\hat{\beta}_1 > \hat{\beta}_2$ at all points in the season. These results support sustained player improvement over season-long streakiness.

Furthermore, the average age of the steady starter sample is much higher than league average (30.2 years vs. 28.2 years) [Eddy, 2021]. An experienced veteran is less likely to substantially improve over the duration of a single season because he has already received ample playing time.

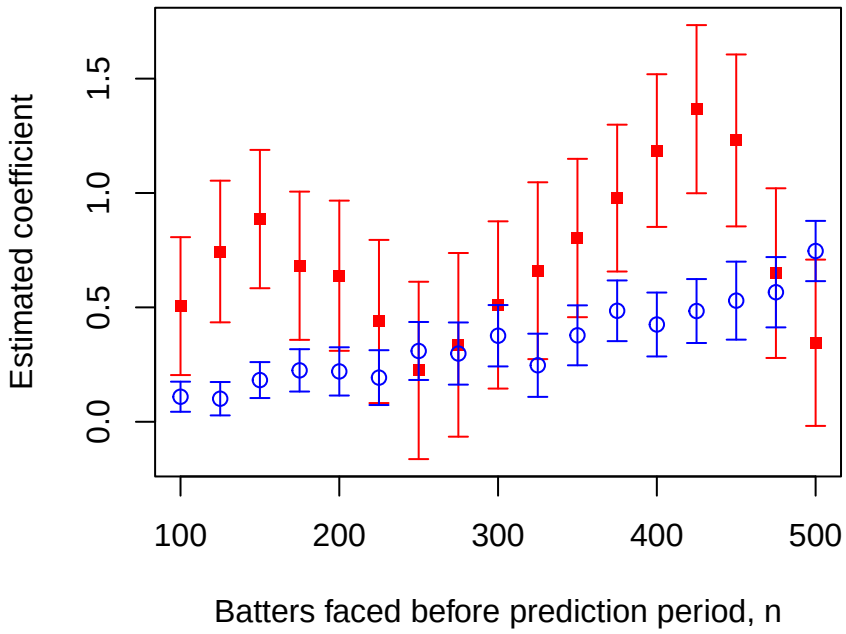

Figure 8: Restricted pitcher sample: estimated coefficients of *current* (as blue circles) and *last3* (as red squares) in eq. (1). Error bars represent heteroskedastic-robust standard errors.

Among pitchers, *last3* follows a trend line that varies greatly from that shown in Figure 3: $\hat{\beta}_1$ has large standard errors and fluctuates significantly. Figure 8 still shows a decrease in $\hat{\beta}_1$ over time while $\hat{\beta}_2$ follows a similar trajectory to that shown in Figure 3, however.

4.5 Isolating current performance

In addition to comparing the predictiveness of *current* and *last3*, I am interested in determining whether *current* adds any additional information to a manager's decision. More specifically, I aim to determine whether a player's current season is particularly predictive

of his subsequent performance, given all available information. Therefore, I estimate an adjusted version of eq. (1) where I replace *last3* with a variable *combine*. *combine* measures a player's average performance over the *last3* and *current* periods. For example, in the hitter sample corresponding to $(500, 25, 100)$, *combine* measures a hitter's average wRC+ over his previous three seasons and the first 500 PA of the current season. By including *current* as an explanatory variable, I aim to evaluate whether a player's current season performance is particularly predictive after considering the player's whole past performance. I show the corresponding results in Figure 9 and Figure 10.

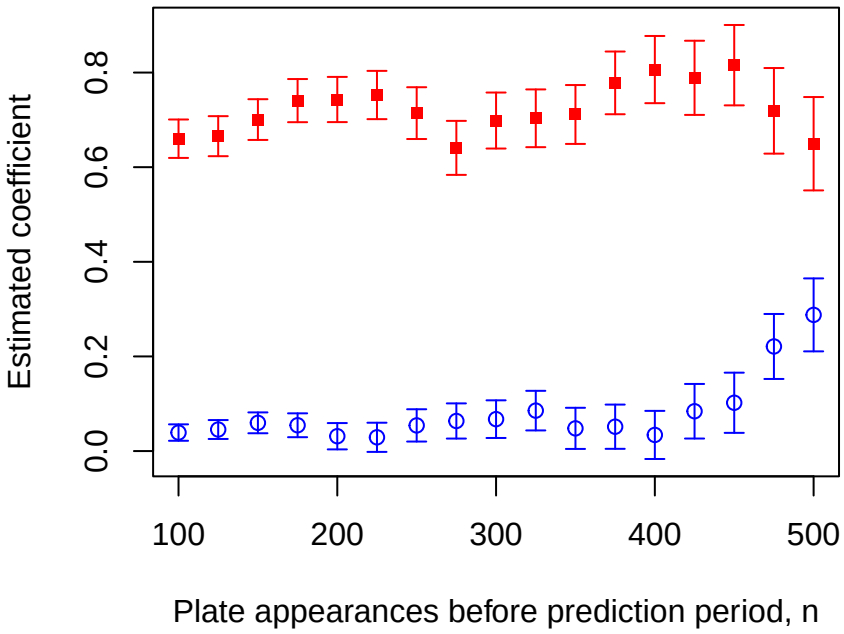

Figure 9: Hitters: estimated coefficients of *combine* (as red squares) and *current* (as blue circles)in eq. (1) where *combine* replaces *last3*, by plate appearances. Error bars represent heteroskedastic-robust standard errors.

The trends shown in Figure 9 and Figure 10 are remarkably similar to those in Figure 1 and Figure 3, respectively. Among hitters, *current* does not add additional predictive value until after the 400 PA mark. After this point, its estimated coefficient becomes significantly positive. I find that among pitchers, *current* provides predictive value from the start. When

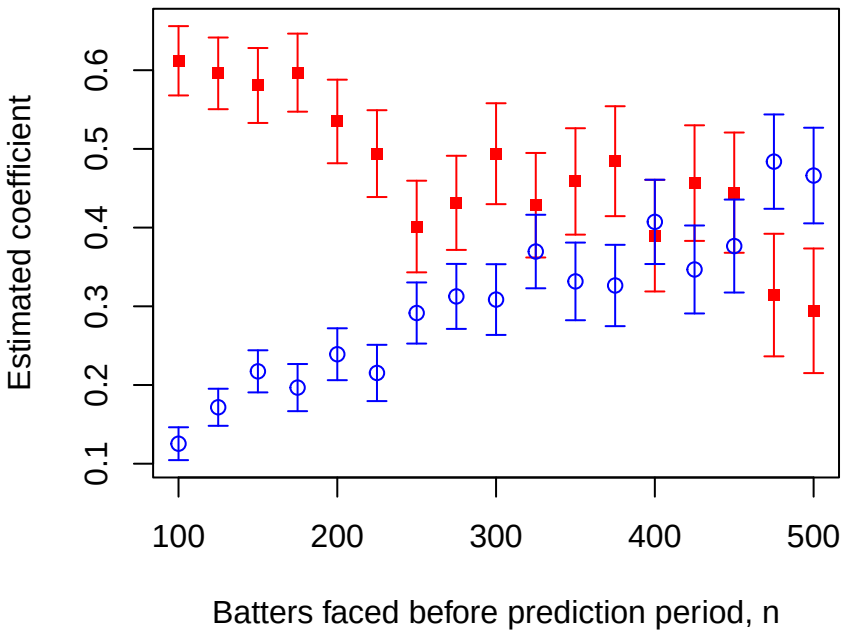

Figure 10: Pitchers: estimated coefficients of *combine* (as red squares) and *current* (as blue circles)in eq. (1) where *combine* replaces *last3*, by batters faced. Error bars represent heteroskedastic-robust standard errors.

$n = 100$ BF, *current*'s estimated coefficient is significantly positive. Consistent with Figure 3, the estimated coefficient corresponding to *current* rises and the estimated coefficient corresponding to *combine* decreases (as with *last3* in Figure 3) as the number of batters faced in the current season increases.

4.6 Heterogeneity

I also group players by various characteristics to estimate eq. (1) using these subsets for sample selection. I consider pools of older veterans and younger players. I also categorize pitchers as strikeout pitchers or contact pitchers and starters or relievers. A pitcher is a strikeout pitcher if he averages more career strikeouts per nine innings than the MLB league average. Hitters are categorized by position and as either contact hitters or power hitters. I

define a power hitter to be a player who averages a career isolated power (ISO) value greater than the MLB league average, where ISO = SLG − BA, the difference between a player's slugging percentage and his batting average. Pitchers who are not strikeout pitchers are characterized as contact pitchers and hitters who are not power hitters are characterized as contact hitters. Despite regression results closely reflecting overall results, there are a few informative exceptions. For power hitters, current season performance holds more weight than it does in overall and contact hitter samples by the end of the season. Furthermore, a higher percent of the variance in *next* is explained by the independent variables on the power hitter sample as compared to the overall hitter sample ($R^2 = 0.306$ vs. $R^2 = 0.2438$).

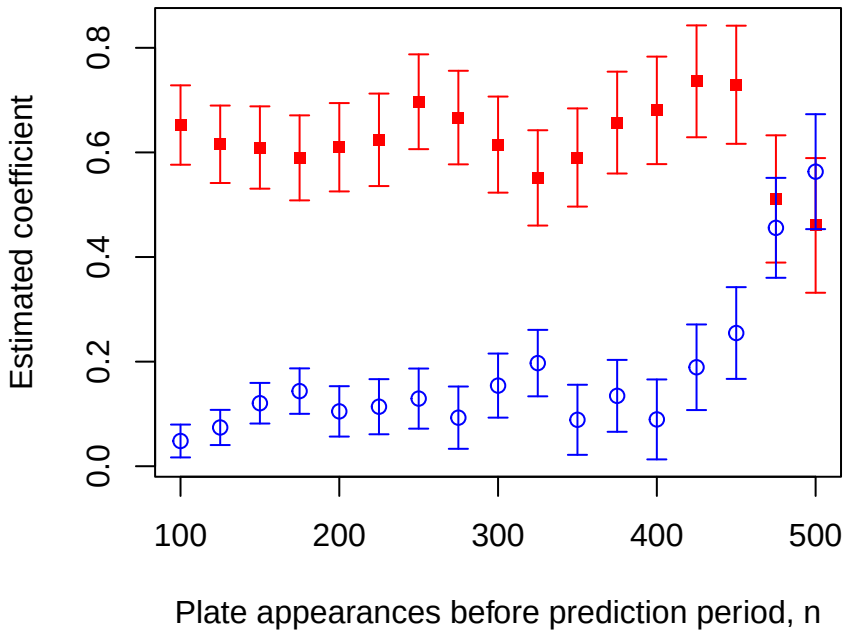

Figure 11: Power hitters: estimated coefficients of *last3* (as red squares) and *current* (as blue circles)in eq. (1) by plate appearances. Error bars represent heteroskedastic-robust standard errors.

For strikeout pitchers, a similar effect arises: current season performance holds more weight at the end of the season than it does in the overall and contact hitter samples. Figure 12 demonstrates this phenomenon. Compared to Figure 3, there are similar trajec-

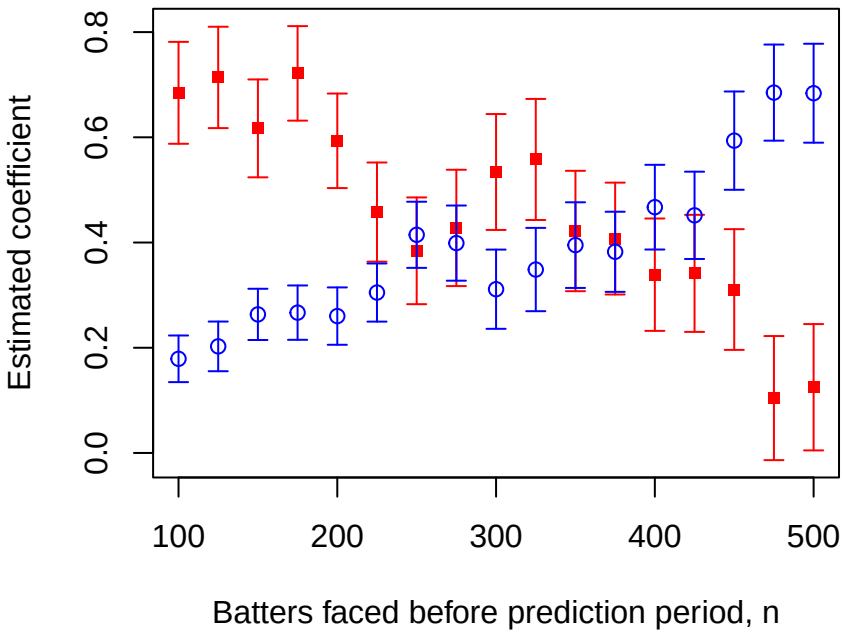

Figure 12: Strikeout pitchers: estimated coefficients of *last3* (as red squares) and *current* (as blue circles)in eq. (1) by plate appearances. Error bars represent heteroskedastic-robust standard errors.

tories corresponding to *last3* and *current*, but *last3*'s trajectory dips lower in Figure 12 than in Figure 3. Similarly, *current*'s trajectory rises higher in Figure 12 than in Figure 3. When $n = 500$ BF, the independent variables explain 42.52 percent of the variance in *next* ($R^2 = 0.4252$), a significant increase from $R^2 = 0.3494$ evident in the full pitcher sample population regression.[5]

These results may inform beliefs concerning the true predictive value of *current* and *recent* performance. Both power hitter and strikeout pitcher performance outcomes are less susceptible to chance than the average player performance. Putting a ball into play allows for uncertainty in the play's outcome; when a ball is put in play, the defense determines the play's outcome. If a pitcher gets batters out by striking out a large proportion of his

[5]I refrain from displaying results where BF $\in \{525, 550\}$ because the samples collected in these settings contain few player observations and the estimates are imprecise.

opponents, the defense has less opportunities to add noise to measures of performance. For hitters, a high ISO value reflects a high home run rate. Similarly, higher home run rates result in less balls that the defense fields (compared to lower home run rates). Therefore, I expect *current* and *last3* to be more predictive of subsequent performance within these subsets, compared to the overall player population. R^2 values reflect this expectation. Since *current* holds more weight within these subsets of players, these types of players may be more susceptible to performance fluctuations between seasons performance and less susceptible to midseason performance fluctuations.

5 Manager decision-making

I focus on manager decision-making in the postseason to capture scenarios where winning in the short-term is of primary importance. As mentioned in section 1, I assume that in the postseason, the manager's aims to maximize the probability of victory in the current game. To do so, I assume that the manager should start the players that give his team the best chance of winning.

Hence, I use postseason data to evaluate manager decision-making. In each model, an observation consists of variables containing information about a player's regular season and prior season performances (i.e. *current*, *last3*, and *recent*) obtained from the regular season data described in section 2. Model observations also include variables that rely on the postseason data described in section 2.

5.1 Pitchers

First, I analyze manager decision-making with respect to starting pitchers. Compared to hitters, decisions concerning pitchers are easier to understand. A manager must choose an optimal starter from a pool of options in the first game of a playoff series. To evaluate manager decision-making optimality, I must first differentiate optimal and suboptimal decisions. First, I predict player postseason performance for each possible starter at using

regression from my estimations of eq. (1). For a player to qualify for inclusion in the pool of potential starters, he must have sufficient rest and be of good health. I define a player to be healthy if he plays at all in the postseason[6] and define a player to have sufficient rest if he has not played for the five consecutive days directly preceding the first game of the playoff series. I only examine scenarios in which the manager chooses between (at least) two players with ample amounts of previous playing time. For a player to qualify for a pool, he must have recorded 600 BF in the current season and 324 BF in the previous season. Such a high current season threshold aims to eliminate players returning from serious injury and ensures the manager has significant information from current and previous seasons to make his decision.

Upon filtering players by the aforementioned thresholds, I group players into platoons by team and season. To summarize, each observation in eq. (2) represents the scenario in which a manager has multiple options to choose from to start game one of a playoff series. I eliminate platoons from the sample if the manager's choice of starter does not qualify for the sample. Such elimination may occur if a manager starts a player returning from injury, a player who has not received sufficient rest, or a rookie. In platoons with two or more players, the model predictions determine the optimal starter. The optimal starter is the player from the platoon with the lowest predicted xFIP. All other players are considered suboptimal starters. A platoon of size N includes $N-1$ observations when estimating eq. (2): I compare the single optimal starter with each of the $N-1$ suboptimal starters. I compare the manager's decision to the model optimal decision and categorize the manager's decision as a *mistake* if the decisions differ. I estimate the coefficients in eq. (2) using OLS regression.

$$mistake = \beta_0 + \beta_1 dlast3 + \beta_2 dcurr + \beta_3 drecent + \gamma C + \mu \tag{2}$$

where

[6]This assumption is not perfect. However, I assume a player to be healthy if he meets the strenuous sample selection criteria. If a player records at least 600 BF in the regular season and a nonzero amount of playing time in the postseason, he is likely to be healthy.

- *mistake* is a dummy variable. *mistake* = 1 if a manager makes a mistake and 0 otherwise.
- *dlast3* = suboptimal starter *last3* - optimal starter *last3*, as measured by xFIP
- *dcurr* = suboptimal starter *current* - optimal starter *current*, as measured by xFIP
- *drecent* = suboptimal starter *recent* - optimal starter *recent*, as measured by xFIP
- C is a vector of controls that includes year and opponent fixed effects
- μ: error term

I use robust standard errors. Note that for *dcurr*, *dlast3*, and *drecent*, a positive value implies that over the respective time period, the suboptimal player records a higher xFIP than the model optimal starter. Since a higher xFIP reflects worse performance than a lower one, a positive difference implies that the optimal starter played better in the specified time frame. A negative difference implies that the suboptimal player played better over the same period. Consider a negative coefficient β_i on $x \in \{$*dcurr, dlast3, drecent*$\}$. Then, when x is negative (the suboptimal player has played better over the specified time period), $\beta_i x > 0$. Hence, the manager is more likely to make a mistake when the suboptimal player has played better over the specified time period. When the optimal player outperforms the suboptimal player over the time period defined by x, x will be positive and $\beta_i x < 0$. Hence, a manager is less likely to make a mistake when the optimal player outperforms the suboptimal player over the time period specified by x. Now, consider a positive coefficient β_i on $x \in \{$*dcurr, dlast3, drecent*$\}$. When $x > 0$, $\beta_i x > 0$. In this scenario, a manager is more likely to make a mistake when the suboptimal player has played worse than the optimal player in the specified time period. Such a scenario represents an underreaction to player performance in the specified time period. β_i may be positive if there are severe overreactions to other variables. I show the corresponding results in Table 3.

Note that the estimated coefficient on *drecent*, $\hat{\beta}_3$, is significantly negative. As I explain above, a negative estimated coefficient implies that a manager is less likely to make a mistake

Table 3: Evaluation of decision-making with respect to pitchers: OLS regression results.

	OLS with Robust SE
	(1)
dlast3	−0.043
	(0.081)
drecent	−0.124***
	(0.026)
dcurr	−0.059
	(0.084)
Observations	170
Adjusted R^2	0.156
Note:	*p<0.1; **p<0.05; ***p<0.01

when the optimal player has performed better recently. On the other hand, the manager is more likely to start the suboptimal starter when the optimal starter has played poorly relative to the suboptimal player recently. This result implies that managers overreact to recent performance, exhibiting hot hand bias.

After finding that managers overreact to recent performance, I separate the optimal starter's statistics and the suboptimal starter's statistics into distinct variables to estimate their effects on the manager's overreaction. When the manager overreacts to differences in recent performance, is he reacting to the suboptimal starter's stellar play, the optimal starter's poor play, or both? I use the same sample that I use to generate results in Table 3. Instead of choosing *dcurr*, *drecent*, and *dlast3* as explanatory variables, however, I use:

- *ocurr*: the optimal starter's *current* value
- *scurr*: the suboptimal starter's *current* value
- *orecent*: the optimal starter's *recent* value
- *srecent*: the suboptimal starter's *recent* value
- *olast3*: the optimal starter's *last3* value

- *slast3*: the suboptimal starter's *last3* value

to quantify the estimated effect of each player's performance across different time periods on the likelihood of a manager mistake. I show the results in Table 4.

Table 4: Estimated coefficients of eq. (2) where each difference variable is split by optimal and suboptimal starter. Dependent variable: *mistake*.

	OLS with Robust SE
olast3	0.158
	(0.100)
slast3	0.065
	(0.104)
orecent	0.121^{***}
	(0.046)
srecent	-0.136^{***}
	(0.035)
ocurr	0.028
	(0.097)
scurr	−0.080
	(0.103)
Observations	170
Adjusted R^2	0.161
Note:	*p<0.1; **p<0.05; ***p<0.01

I find an overreaction to recent performance consistent with my findings from Table 3. The estimated effects of the suboptimal and optimal starter's recent performance are nearly identical in magnitude. I confirm this observation by running an F-test and find an insignificant difference ($p = 0.81$) between coefficient magnitudes. Hence, managers react similarly to positive recent play by the suboptimal starter and poor recent play by the optimal starter when making decisions.

5.2 Hitters

I evaluate manager decision-making with respect to hitters in determining when, between games, a manager should bench a starter in favor of another player at the same position. I consider postseason performance for identical reasons expressed in the previous section. My

setup is as follows. The manager starts a player at each position in Game 1 of a playoff series. In each subsequent game of the series, the manager has multiple players that he may choose to start. From each pool, he can only start one per position. For each player,estimations of eq. (1) predict his subsequent performance based on his PA totals performance (as they did in the previous section). The model determines the optimal player starter to be the one with the highest predicted wRC+. From game data, I observe the true starter the manager chose to play. Comparing the model's choice to the manager's decision, I consider the manager's decision a *mistake* if the two results differ. These mistakes can be categorized as one of two kinds of mistakes.

1. The manager should have started the previous starter in the next game but benched him in favor for a suboptimal starter.

2. The manager should have benched the previous starter in favor for the model optimal starter (but did not).

I estimate eq. (3) to analyze manager decision-making tendencies with respect to player performance across different time periods.

$$\begin{aligned} mistake = \beta_0 + \beta_1 current + \beta_2 last3 + \beta_3 recent + \beta_4 scurrent + \\ \beta_5 slast3 + \beta_6 srecent + \beta_7 post + \gamma C + \mu \end{aligned} \tag{3}$$

where *current*, *recent*, and *last3* are the previous starter's statistics (defined in section 3) and *scurrent*, *srecent*, and *slast3* are the proposed substitute's statistics (defined in section 3). I define *post* to be the starter's postseason performance up until the game before the starter gets pulled, or up until the last game in the playoff series in the case where the starter starts the entire series. More specifically, *post* is a player's weighted on-base average (WOBA), defined by Slowinski [2010b]. C is a control vector including year fixed effects.

I evaluate eq. (3) for each game in a playoff series, starting with the second game. I use playoff series data from $2006-2021$ to form platoons. I only consider platoons in which both

the game one starter and his potential substitute reach at least 324 PA over the previous three seasons and at least 125 PA in the current season to predict each player's subsequent performance from eq. (1)'s estimations. At each game, the manager either pulls the previous game's starter or he does not. If he does not pull the starter at all, then the starter starts each game of the entire series. Given the two forms of mistakes possible, I partition the set of decisions into two:

- A set where eq. (1) recommends not pulling the starter
- A set where eq. (1) recommends pulling the starter

First I estimate eq. (3) with decisions where the manager shouldn't pull the starter. Here, a *mistake* is defined to be a scenario in which the manager pulls the starter in favor of the suboptimal substitute. Hence, there is at least one game in the playoff series where the optimal starter does not start. I use robust standard errors and show the results in Table 5.

Note that wRC+ measures performance in the opposite direction of the pitching metric, xFIP. For pitchers, recall that a better xFIP is a lower one; for hitters, a better wRC+ is a higher one. Rather than using differences in player statistics as explanatory variables in equation 3, as done with pitchers, I treat both the starter and substitute's statistics as distinct explanatory variables. The asymmetry of available information between substitutes and starting players causes me to distinguish between starter and substitute statistics. Assuming good health, it is reasonable to assume similar playing time between two starting pitchers. Among hitters, since players do not typically operate as part of a regularly rotating platoon, postseason playoff starters commonly obtain more regular-season plate appearances than substitutes.

current and *recent*, the statistics of the previous game's starter, have statistically significant and negative coefficients. Hence, the better the previous starter plays, the less likely a manager is to prematurely bench a player for the suboptimal substitute. The estimated coefficient on *post* is insignificant. This result may imply that managers react adequately to very recent performance. However, the estimated coefficient on *recent* is significantly neg-

Table 5: Evaluation of decision-making with respect to managers pulling starting hitters: OLS regression results. Dependent variable = *mistake*, pulling the hitter earlier than optimal.

	OLS with Robust SE
	(1)
current	-0.418^{**}
	(0.182)
recent	-0.173^{***}
	(0.061)
last3	0.081
	(0.221)
srecent	0.137
	(0.085)
scurrent	0.156
	(0.181)
slast3	0.154
	(0.289)
post	−0.049
	(0.079)
Observations	165
Adjusted R^2	0.191
Note:	*p<0.1; **p<0.05; ***p<0.01

ative and the estimated coefficient on *srecent* is significantly positive. These results imply that a manager is more likely to prematurely pull a starter both when the starter has played poorly and when the substitute has played well over recently. Hence, these results imply that managers overreact to recent performance and are consistent my findings concerning pitching decisions[7].

In the second scenario, a *mistake* is defined as the decision to start the previous starter in the next game when he should be pulled in favor of the substitute. I maintain one observation per platoon by considering decisions by series. If the starter plays the entire series, a decision is considered a mistake. Otherwise, the manager correctly pulls the starter at some point in the playoff series. I regress eq. (3) on decisions where the model suggests benching the starter. Once again, I use robust standard errors, include year fixed effects and display the results in Table 6[8].

The refined pool includes few model observations, yielding large standard errors. Despite this, the estimated coefficient on *current* deviates significantly from zero. Hence, I interpret the manager to be more likely to make a mistake when the current game starter has been playing well in the current season. Given that a player is playing well in the current season, the manager continues to start the starter for longer than optimal. Similarly, the estimated coefficient on *post* is significantly positive. Managers are expected to be more likely to keep a starter in the series longer than optimal when he has played well in the postseason thus far. The positive coefficient on *post* may imply that managers overreact to postseason performance in addition to recent performance. Managers may weigh postseason performance too greatly for multiple reasons, despite small sample sizes. For example, managers may assign clutch factors to certain players who perform well in the early postseason and believe that players who are clutch are more likely to continue playing well in the postseason than those who are not clutch.

[7]Of the 165 decisions included in the sample, the model classified 41 as mistakes.
[8]48 of the 83 decisions in this decision pool are considered mistakes.

Table 6: Evaluation of decision-making with respect to hitters: OLS regression results. Dependent variable = *mistake*, pulling a starting hitter later than optimal.

	OLS with Robust SE
	(1)
current	0.534
	(0.324)
recent	-0.230^{*}
	(0.124)
last3	-0.107
	(0.605)
srecent	-0.119
	(0.176)
scurrent	0.509
	(0.317)
slast3	-0.417
	(0.357)
post	0.363^{***}
	(0.134)
Observations	83
Adjusted R^2	0.140
Note:	*p<0.1; **p<0.05; ***p<0.01

5.3 Redefined mistake

In each of the two previous subsections, my methodology deems a manager's decision to be a *mistake* if an unselected player is predicted to outplay the chosen starter. The margin for error here is small: if the unselected player is predicted to outplay the selected starter by even an infinitesimally small amount, the methodology classifies this decision as a *mistake*. Predictions of future performance are obviously noisy. The explanatory variables included in eq. (1) only explain a fraction of the variance in player performance. Model predictions are based on a subset of information available to the manager which may prove insignificant relative to additional available information not captured by eq. (1). Managers have access to both private information and information not reflected by previous player performance. It follows that their decisions are not based exclusively on a weighting of a player's *current* season performance, *recent* performance, prior season performance and captured controls. Therefore, I use a wider margin of error to evaluate whether a manager acts optimally.

I redefine a *mistake* to be a decision which differs substantially from the model optimal choice. A decision is defined to be a *mistake* if the predicted performance of the optimal player differs from the predicted performance of the suboptimal player by at least one standard deviation. For pitchers, a decision is considered a *mistake* if the manager-chosen starter has a predicted xFIP at least one standard deviation higher than that of the model-chosen starter. For hitters, a decision is considered a *mistake* if the manager-chosen starter has a predicted wRC+ at least one standard deviation lower than that of the model-chosen starter. Once again, I split decisions about hitters into two decision pools:

- situations where the manager should bench the previous game starter, and
- situations where the manager should not bench the previous game starter.

Mistakes are categorized as before. I estimate eq. (2) and eq. (3) using the updated pools.

I display the regression results obtained from the sample of pitching decisions using the refined categorization in Table 7. As before, the estimated coefficient on *drecent* is negative, further solidifying the claim that managers overreact to recent performance.

Table 7: Pitching decisions: estimated coefficients of eq. (2) where each difference variable is split by optimal and suboptimal starter. Dependent variable: *mistake* (the redefined, rigorous version).

	OLS with Robust SE
	(1)
dlast3	−0.108
	(0.086)
drecent	−0.100***
	(0.029)
dcurr	−0.035
	(0.087)
Observations	156
Adjusted R^2	0.095
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 8: Refined hitter decisions: evaluation of decision-making with respect to managers pulling starting hitters: OLS regression results. Dependent variable = *mistake*, pulling the hitter earlier than optimal.

	OLS with Robust SE
	(1)
current	−0.410**
	(0.159)
recent	−0.107*
	(0.060)
last3	0.061
	(0.190)
srecent	0.096
	(0.072)
scurrent	0.104
	(0.148)
slast3	0.033
	(0.244)
post	−0.067
	(0.088)
Observations	219
Adjusted R^2	0.141
Note:	*p<0.1; **p<0.05; ***p<0.01

I display regression results obtained from the samples of hitting decisions using the refined categorization in Table 8. Note that the sample sizes of each regression have changed. Recall that decisions are grouped by whether the model determines that a starter should be benched. By redefining the criteria that determines when a player should be benched, benching threshold increases in rigor. Hence, the model suggests that 219 players should not be benched and 29 should be benched. Previously these suggestions were 165 and 83, respectively. The pool with a mere 29 model observations does not yield significant results[9]. However, when the model suggests pulling the starter, my estimates are more precise. As found originally, *recent* maintains a negative estimated coefficient. These results solidify the claim that managers overreact to recent performance. Despite finding evidence for overreactions to recent performance (which may signal implicit underreaction to other performance periods), I do not find managers explicitly underreacting to previous player performance over any of the defined time periods in any of the aforementioned scenarios and regressions. While I find significant overreaction to recent performance in manager decision-making with respect to hitters and pitchers, it is admittedly difficult to evaluate manager performance based on limited numbers of player observations and incomplete information.

6 Conclusion

Managers should value current season performance more for pitchers than hitters when comparing current season performance to a player's performance over his previous three seasons. This finding is consistent throughout the season. A player's performance in the current season increases in predictiveness over time from the 100 BF mark through his first 500 BF. Among hitters, current season performance carries roughly the same predictive value through a player's first 400 PA before increasing relative to his prior performance. By the end of the season, a pitcher's current season performance is significantly more predictive of subsequent performance than his performance over his previous three seasons. This is not the

[9]The small sample size likely created significantly large estimated standard errors.

case for hitters. Among hitters, I find no point in the regular season where a player's current season performance is more predictive of subsequent performance than his performance over the previous three seasons. Hence, I conclude that future pitching performance is more dependent on current season performance relative to hitters when predicting performance over a single season's duration.

Discovering that adding future season performance as a control variable does not significantly change estimated coefficients on *current* and *last3* further justifies this conclusion. In fact, upon adding future performance as a control, the high estimated coefficients on current season performance and near-zero coefficients on a player's performance in his prior three seasons found at the end of the regular season may imply that for pitchers who do not improve, current performance becomes an even more important indicator by the end of the season. Significant current season predictive value, upon controlling for future performance, provides key evidence of a season-long hot hand. If a player is performing at a certain level in the current season because this performance level accurately reflects his baseline ability, such a performance level (and baseline ability) should be maintained in future seasons. Hence, by controlling for future season performance, I aim to isolate a season-long hot hand effect in the current season. Such a control is not perfect, however: if a player improves (or worsens) in baseline ability from season to season, his future season performance will fail to perfectly capture his current season baseline ability. More evidence for a season-long hot hand surfaces when I examine subsets of players. Current season performance is particularly predictive for power hitters and strikeout pitchers, players with more control over their performances. This finding is consistent with that of of Descamps et al. [2022]. A player's initial success (or failure) in the current season is likely to affect a player's subsequent performance. These effects may be more visible among players with more control over their measures of performance.

With respect to manager choice, When aiming to maximize a team's probability of immediate victory, managers value recent pitcher play strongly. This phenomenon holds both when managers are determining which pitcher to start out of a pool of potential starters and when deciding how long the starting pitcher should continue playing within a game (see ap-

pendix A.1 for details). The stronger a pitcher has played recently, the longer I expect a manager to keep the starter in the game after controlling for performance within the game. I characterize this behavior as an overreaction to the hot hand. Out of two players who perform similarly, managers tend to favor the one who has played better recently. Furthermore, I find that when choosing a starting pitcher, managers tend to overvalue a player's recent performance. Holding all else constant, managers are less likely to start a player who has performed poorly at the end of the regular season and more likely to start a player playing well. While the estimation of eq. (1) demonstrates that the manager should value recent performance to some degree[10], I find that managers overvalue such recent performance. These short-term overreactions are consistent with the literature [Bordalo et al., 2019, Wang, 2021] and my hypotheses. These results provide additional evidence of short-term overreactions by actors in high-stakes forecasting scenarios.

Managers tend to value a hitter's current season performance strongly. The estimations of eq. (1) and its corresponding predictions of subsequent hitting and pitching performance causes my expectations to differ from the manager's actions. Estimations of eq. (1) made when players have high amounts of playing time (PA, BF $>$ 500) imply that managers should value current season performance more for hitters than for pitchers relative to prior career performance. However, the weights placed on hitters may be more symmetric than optimal. Managers appear to overvalue hitter current season performance when selecting an optimal starter out of a candidate pool. As with pitchers, managers tend to overreact to the hot hand. I find that managers have a shorter leash on starters than what eq. (1) considers optimal when the substitute player has played well lately and the starter has not. The manager is more likely to pull a starting hitter prematurely when he has played poorly recently. Moreover, the manager is more likely to pull a starting hitter prematurely when the substitute player has played well recently. While general microeconomic theory implies that as a substitute good increases in value, it becomes more appealing relative to the original good, I find that managers over-update their beliefs with respect to recent performance.

[10]Refer to appendix A.3 for estimated *recent* coefficients.

I find no anchoring bias with respect to prior season performance within manager decision-making tendencies. Specifically, I find no scenarios where the manager overvalues a player's prior season performance. However, my aforementioned finding concerning manager over-reaction to current season performance may indicate a current season anchoring bias; if I consider the start of the current regular season to be an opportunity for managers to re-set their beliefs concerning player ability to some degree, as is consistent with Dai's [2018] findings, then perhaps the overvaluation of current season performance that a manager exhibits when he pulls the starter later than optimal indicates anchoring bias. However, the positive estimated coefficient on current season performance loses significance when I refine my definition of a manager *mistake*. Therefore, I cannot conclude that managers overvalue current performance in this scenario. Alternatively, I find an underreaction to current season performance when a manager decides to pull a starting hitter prematurely (although my sample size is small). Upon increasing the threshold for eq. (3) to consider a manager's decision to be a *mistake*, the negative estimated coefficient on current season performance loses significance. Such inconsistent findings may imply that actors find it difficult to properly weigh performance across multiple time periods when making decisions concerning future performance.

6.1 Future work and extensions

Future work may include estimating eq. (1) using alternative criteria for sample selection. The criteria for player inclusion in these samples may allow for varying the threshold criteria for previous player performance. For example, investigators may obtain a measure for baseline player ability by measuring prior performance over one season instead of three. Additional studies may partition the current season into discrete time periods of which managers may assign predictive value rather than looking at current season performance as a single entity (after controlling for recent performance). Additional data collection of post-season performance and decisions made regarding hitters may provide more insight into how managers react to hitter performance across various time periods. Rather than using contin-

uous measures of performance, categorizing players into states based on recent performance, such as hot and cold, may inform decision-making on a granular level. One may obtain estimates of weights for periods of hot play and cold play separately rather than using one estimate when determining how to weigh recent performance when predicting subsequent performance. One may naturally extend this work by searching for a a hot-hand over a wider time horizon than I do. Given performance in multiple distinct seasons, how much weight should be given to each season when predicting player performance in subsequent season (or seasons)? General managers must consider these factors when paying players to join their teams. Admittedly, decisions here should most likely be taken on a case-by-case basis. However, assigning weights to each season may provide a useful baseline estimate of expected future performance.

References

1990 standard batting. *Baseball-Reference.com*, a.

1991 standard batting. *Baseball-Reference.com*, b.

P. Ayton and I. Fischer. The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & cognition*, 32(8):1369–1378, 2004.

D. J. Benjamin. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2:69–186, 2019.

P. Bordalo, N. Gennaioli, R. L. Porta, and A. Shleifer. Diagnostic expectations and stock returns. *The Journal of Finance*, 74(6):2839–2874, 2019.

H. Dai. A double-edged sword: How and why resetting performance metrics affects motivation and performance. *Organizational Behavior and Human Decision Processes*, 148: 12–29, 2018.

A. Descamps, C. Ke, and L. Page. How success breeds success. *Quantitative Economics*, 13 (1):355–385, 2022.

M. Eddy. The shape of baseball is changing in 2021. 2021.

T. Gilovich, R. Vallone, and A. Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, 17(3):295–314, 1985.

B. Green and J. Zwiebel. The hot-hand fallacy: Cognitive mistakes or equilibrium adjustments? evidence from major league baseball. *Management Science*, 64(11):5315–5348, 2018.

J. K. Hakes and C. Turner. Pay, productivity and aging in major league baseball. *Journal of Productivity Analysis*, 35(1):61–74, 2011.

J. P. Jamieson. The home field advantage in athletics: A meta-analysis. *Journal of Applied Social Psychology*, 40(7):1819–1848, 2010.

W. Leitch. 30 reasonable goals – 1 for each team. *MLB*, 2022.

J. B. Miller and A. Sanjurjo. Surprised by the hot hand fallacy? a truth in the law of small numbers. *Econometrica*, 86(6):2019–2047, 2018.

J. B. Miller and A. Sanjurjo. A bridge from monty hall to the hot hand: The principle of restricted choice. *Journal of Economic Perspectives*, 33(3):144–62, August 2019. doi: 10.1257/jep.33.3.144. URL https://www.aeaweb.org/articles?id=10.1257/jep.33.3.144.

J. B. Miller, A. Sanjurjo, et al. A cold shower for the hot hand fallacy. Technical report, IGIER working paper, 2014.

T. Offerman and J. Sonnemans. What's causing overreaction? an experimental investigation of recency and the hot-hand effect. *The Scandinavian Journal of Economics*, 106(3):533–554, 2004.

M. Ozanian and J. Teitelbaum. Baseball's most valuable teams 2022: Yankees hit 6 billion as new cba creates new revenue streams. *Forbes*, 2022.

M. Raab and B. Gula. Hot hand belief and hot hand behavior: A comment on koehler and conley. *Journal of sport & exercise psychology*, 26:167–171, 03 2004. doi: 10.1123/jsep.26.1.167.

D. Richards. Going deep: The relative value of fip, xfip and siera. *Pitcherlist*, 2019.

H. Rosenfeld. Iron man: The cal ripken, jr. *Story (New York,[1995] 1996.)*, 4, 1995.

D. Schoenfield. Mlb hitters who are missing their age-27 peak season in 2020. *ESPN*, 2020.

P. Slowinski. wrc and wrc+. *FanGraphs*, 2010a.

P. Slowinski. woba. *FanGraphs*, 2010b.

D. F. Stone. Measurement error and the hot hand. *The American Statistician*, 66(1):61–66, 2012.

D. F. Stone and J. Arkes. March madness? underreaction to hot and cold hands in ncaa basketball. *Economic Inquiry*, 56(3):1724–1747, 2018.

C. Wang. Under-and overreaction in yield curve expectations. *Available at SSRN 3487602*, 2021.

A Appendix

A.1 Manager decision-making: behavior

Before evaluating manager decision-making optimality with respect to player performance over various prior playing periods, I identify a major behavioral trend within manager decision-making. Stone and Arkes [2018] find that decision makers underreact to the short-term hot hand in college basketball. In light of this finding, I evaluate to what degree managers react to short-term streakiness when pulling starting pitchers by using OLS regression on postseason data to estimate eq. (4). Each model observation is one player-game combination instead of a player-season combination as used in eq. (1).

$$\text{totBF} = \beta_0 + \beta_1 last3 + \beta_2 current + \beta_3 recent + \beta_4 inGame + \beta_5 recent \cdot inGame + \gamma C \quad (4)$$

Here, inGame is a measure of pitcher performance in the current game, totBF is the number of batters the pitcher faces in the current game, C is a vector of controls, and *last3*, *current*, and *recent* are defined as before. I include year fixed effects and a pitcher's average batters faced per game in the regular season as controls.

I am interested in β_5, the coefficient on the interaction between *recent* and inGame. β_5 is the coefficient of interest because it contains information about manager reactions to recent player performance. I interpret a negative estimated coefficient to mean that a starting pitcher to receive less playing time for a given single-game performance when he has played poorly recently relative to a player who has been hot recently. My findings are consistent with this theory: a manager is more likely to pull a player for poor in-game performance when he has played poorly recently. This sign suggests that a manager has a shorter leash on players who have been cold recently (high *recent* xFIP) and a longer leash on pitchers who have been hot recently (low *recent* xFIP). I display the regression results in Table 9.

When using robust standard errors, I find that the sign on the interaction term between

Table 9

	OLS with Robust SE
	(1)
inGame	4.101
	(8.355)
recent	1.001
	(0.779)
current	−0.382
	(0.501)
last3	−0.550
	(0.556)
RegBFperGame	0.944***
	(0.158)
factor(game_year)2019	0.405
	(1.056)
factor(game_year)2020	1.675*
	(1.008)
factor(game_year)2021	−1.217
	(1.019)
inGame:*recent*	−5.021**
	(2.191)
Constant	2.931
	(5.653)
Observations	221
Adjusted R^2	0.354
Note:	*p<0.1; **p<0.05; ***p<0.01

inGame and *recent* remains negative but the significance level decreases. Evaluating the same regression model on a sample of middle relievers instead of starters, I find no significant results. Relief pitchers average small numbers of BF per game and I attribute the lack of significant results to this attribute.

A.2 Age

According to ESPN, the average MLB player's peak performance period ends around age 32 [Schoenfield, 2020]. While perhaps not directly related to hot hand analysis, managers may consider the effect of age on performance at various points in the season in conjunction with other information to predict player performance for older (and younger) players. Hence, I estimate equation eq. (1) on the pitching sample, restricting the sample to pitchers aged 32 and older. Since each of these players is older than 27, a player's peak age [Hakes and Turner, 2011], $dfpeak = |age - 27| = age - 27$. Since $dfpeak$ is now a linear transformation of age, I replace $dfpeak$ with age in eq. (1). I am interested in the estimated effect of age on subsequent player performance at various points in the season. I show the estimated coefficients from OLS regression taken across $n \in \{100, 125, ..., 600\}$ in Figure 13.

Note that for BF < 500 the estimated coefficient is positive. Hence, at the beginning of the season, the results may imply that experience boosts player performance. However, the estimated coefficient is negative when BF > 500. These results may imply a fatigue effect for older players relative to the rest of the league in which this fatigue may harm player performance more than experience boosts player performance.

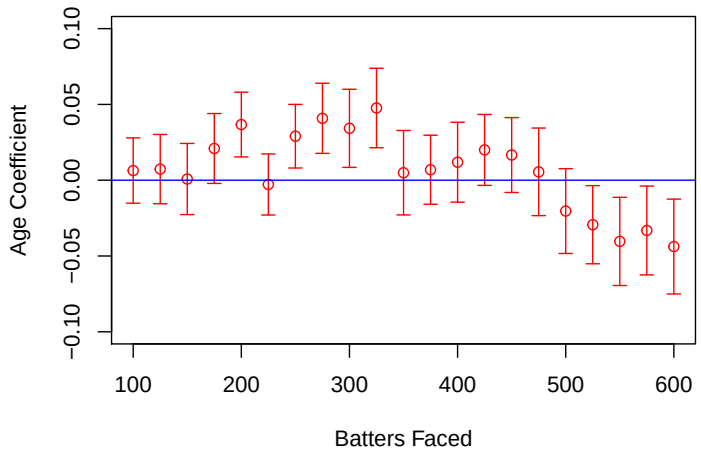

Figure 13: The estimated effect of age on subsequent player performance at various points throughout the season. A positive age coefficient implies that greater player age increases expected subsequent performance, while a negative age coefficient implies that greater player age decreases expected subsequent performance.

A.3 *recent* and *future* estimated coefficients

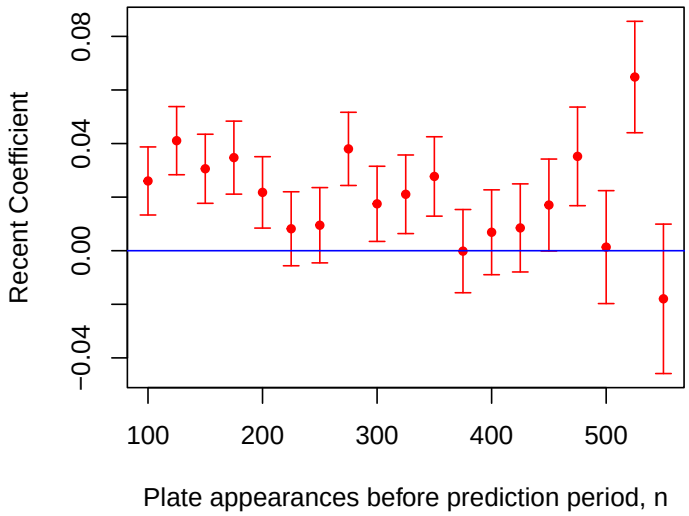

Figure 14: Hitters: estimated coefficients corresponding to *recent* on subsequent player performance from eq. (1) by plate appearance. Error bars represent heteroskedastic-robust standard errors.

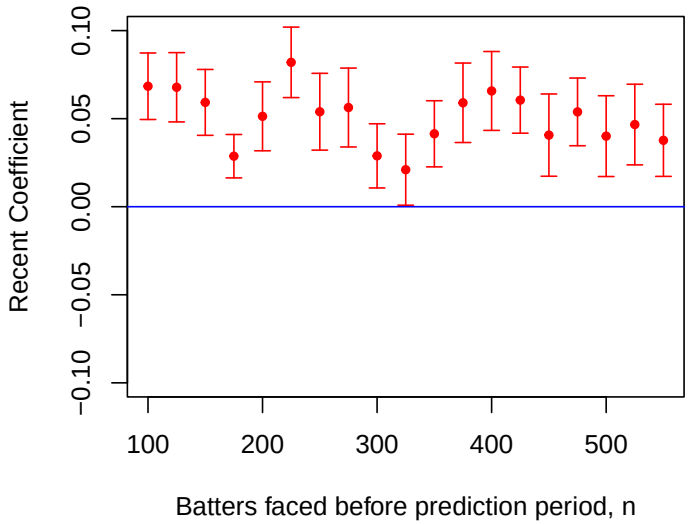

Figure 15: Pitchers: estimated coefficients corresponding to *recent* on subsequent player performance from eq. (1) by batters faced. Error bars represent heteroskedastic-robust standard errors.

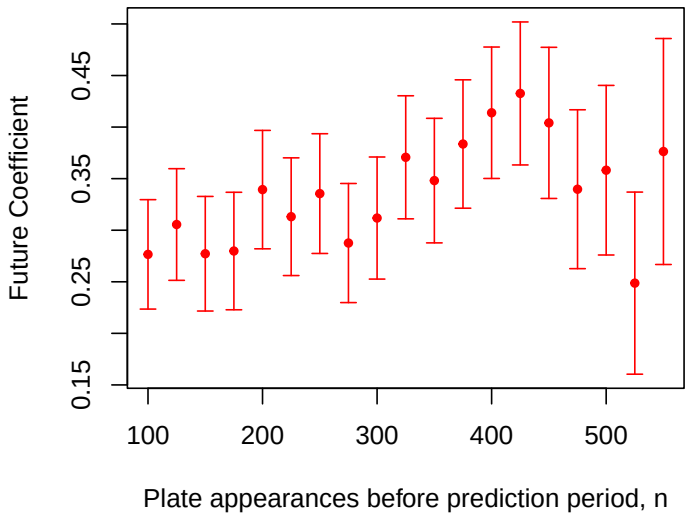

Figure 16: Hitters: estimated coefficients corresponding to *future* on subsequent player performance by plate appearance. Error bars represent heteroskedastic-robust standard errors.

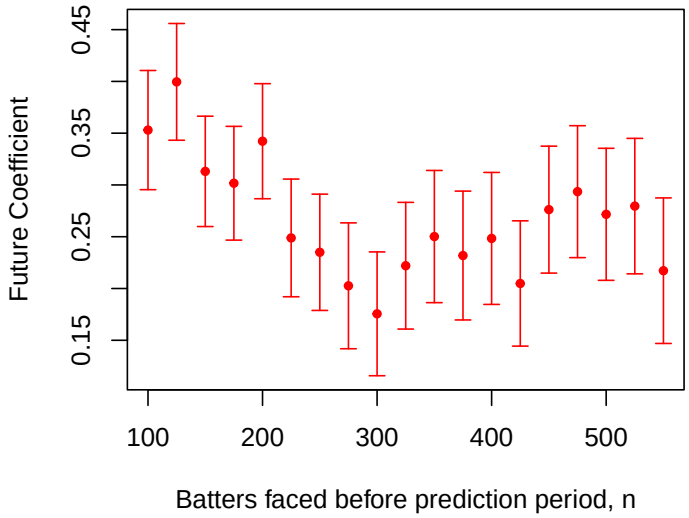

Figure 17: Pitchers: estimated coefficients corresponding to *future* on subsequent player performance by batters faced. Error bars represent heteroskedastic-robust standard errors.

A.4 *last3*: season breakdown

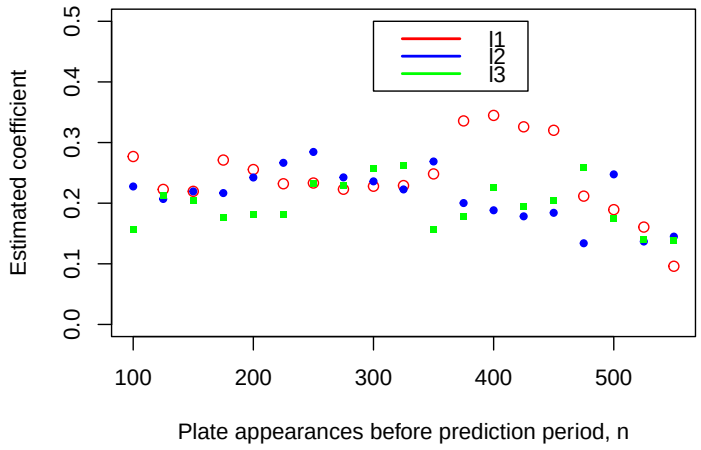

Figure 18: Hitters: estimated coefficients corresponding to l_1, l_2, and l_3 on subsequent player performance by plate appearance. Error bars represent heteroskedastic-robust standard errors.

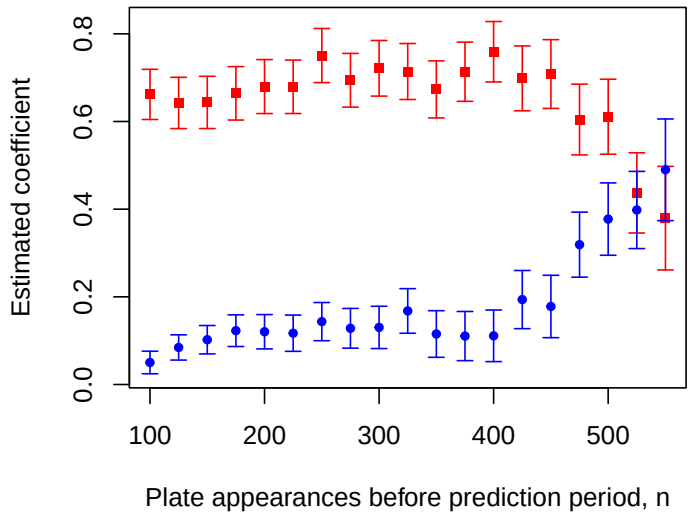

Figure 19: Hitters; *last3* as a weighted average, weighted by season: estimated coefficients of *last3* and *current* on subsequent player performance by plate appearance. Error bars represent heteroskedastic-robust standard errors.

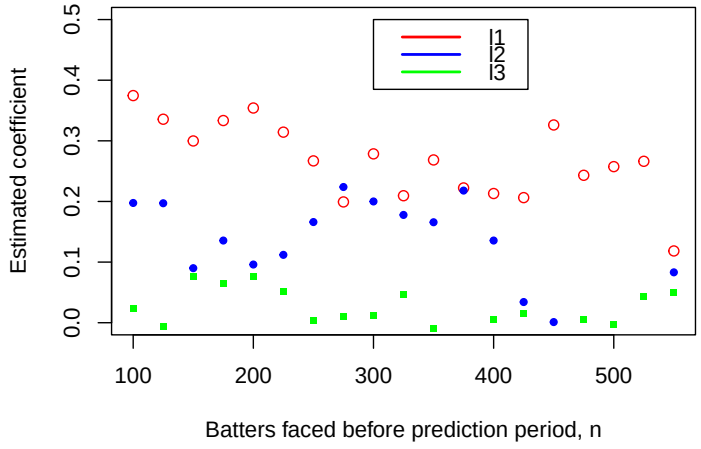

Figure 20: Pitchers: estimated coefficients corresponding to l_1, l_2, and l_3 on subsequent player performance by batters faced. Error bars represent heteroskedastic-robust standard errors.

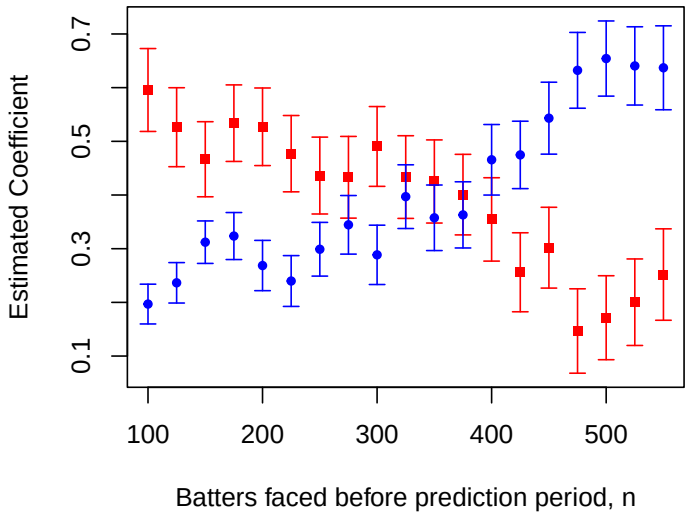

Figure 21: Pitchers; *last3* as a weighted average, weighted by season: estimated coefficients of *last3* and *current* on subsequent player performance by plate appearance. Error bars represent heteroskedastic-robust standard errors.

A.5 Hitter example sample: summary statistics

Table 10: Summary statistics corresponding to the $(550, 25, 100)$ hitter sample. pH = percent home games, oSK = opponent skill.

Statistic	N	Mean	St. Dev.	Min	Max
year	441	2011.036	4.713	2003	2019
last3	441	124.218	20.753	73.920	185.988
last	441	126.044	23.540	68.841	195.672
current	441	126.247	23.888	60.266	208.860
recent	441	127.594	72.693	−50.000	403.880
Next	441	127.216	40.277	22.740	262.825
age	441	29.773	3.004	23	39
dfpeak	441	3.249	2.480	0	12
lastpH	441	0.501	0.014	0.441	0.579
currpH	441	0.501	0.019	0.442	0.552
hotpH	441	0.475	0.333	0.000	1.000
nextpH	441	0.501	0.101	0.227	0.750
futurepH	419	0.497	0.315	0.000	1.000
lastoSK	441	4.205	0.230	3.668	4.600
curroSK	441	4.215	0.232	3.677	4.630
hotoSK	441	4.235	0.290	3.450	4.947
nextoSK	441	4.222	0.251	3.635	4.762
futureoSK	419	4.222	0.299	3.360	5.230
last3oSK	399	4.201	0.214	3.777	4.525
careerISO	441	0.169	0.053	0.000	0.275
careerBA	441	0.259	0.040	0.000	0.314