3-2018

Does Player Performance Outside of Major League Baseball Translate to the MLB?

Ian Vogt

DOES PLAYER PERFORMANCE OUTSIDE OF MAJOR LEAGUE BASEBALL TRANSLATE TO THE MLB?

by

Ian Vogt

* * * * * * * * *

Submitted in Partial Fulfillment
of the requirements for
Honors in the Department of Economics

UNION COLLEGE
March, 2018

ABSTRACT

VOGT, IAN. Does player performance outside of Major League Baseball translate to the MLB? Department of Economics, March 2018.

ADVISOR: Professor Tomas Dvorak

Statistical analysis has transformed the way front offices across Major League Baseball manage the rosters of their teams. However, much of this statistical analysis is limited to evaluating players playing in the American major league environment. Little has been done in the way of using statistical analysis to evaluate how performance translates from league-to-league, and the market for international and college players remains highly inefficient, despite expansion of these player pools. My study is an attempt to make this market a more efficient one.

I measure the correlation between performance in two top international baseball leagues (Nippon Professional Baseball and the Korean Baseball Organization) as well as America's top amateur league (the NCAA) and performance in Major League Baseball. I am studying three performance metrics for both batters and pitchers: strikeout rate, walk rate, and home run rate. I find that these metrics in the foreign and amateur leagues studied account for little of the variation in Major League performance. However, the predictive power of foreign league statistics is not significantly lower than the predictive power of past performance in the MLB, indicating predicting player performance is a difficult task.

TABLE OF CONTENTS

1. Introduction

There is a higher influx of professional players from Korean and Japanese professional leagues into the major leagues than ever before. Before 1990, only two Japanese-born players made it to the major league level. Since then 62 have, with the latest one, Shohei Otani, being one of the top stories from this past offseason. Chan-Ho Park became the first Korean-born player to debut in the major leagues in 1994. In total, only 23 Koreans have played in Major League Baseball, 10 of which have debuted in the past five seasons. Yet even as the market for these players grows, inefficiencies remain. In 2014, the Boston Red Sox thought they had solved their problem in replacing former all-star center fielder Jacoby Ellsbury when they signed Cuban outfielder Rusney Castillo to a 7-year/$72.5 million contract from the Cuban National Series (CNS), the richest contract ever given to a Cuban international free agent. The success of other Cuban professionals who had previously made the transition to the major leagues, such as Yoenis Cespedes, Yasiel Puig, and Jose Abreu, had paved the way for Castillo's payday. However, in the three-plus Major League seasons since signing, Castillo has only appeared in 99 major league games, and was relegated to the Red Sox' top minor league affiliate for the entire 2017 season. Castillo's payday came in the midst of Abreu's rookie year, a season in which Abreu won the American League Rookie of the Year and finished fourth in AL MVP voting. Abreu had signed for 6 years/$68 million the previous winter, a contract that at the time was a record for a Cuban international free agent. Abreu showed the remarkable success a Cuban free agent could have in his first season, expanding the market for Castillo. Yet, despite the richer contract, Castillo's numbers in the CNS were markedly worse than Abreu's. In his last five seasons playing in Cuba,

Abreu never recorded an on-plate plus slugging (OPS) lower than 1.068; Castillo's never went higher than .940 in any of his five seasons playing in the CNS, and he was coming off a season in which his OPS was a meager .770. Was Castillo's performance in the CNS predictive of his MLB performance?

I collect data on individual player statistics from Nippon Professional Baseball (NPB), the Korean Baseball Organization (KBO), and the National Collegiate Athletic Association (NCAA) in order to equate performance in one league to performance in the MLB. I merge this data with individual player data from the MLB and pick out players that have transitioned from a foreign or amateur league to the MLB, measuring how non-MLB performance lines up with MLB performance. I use sample minimums and a set timeframe before and after the transition is made that will be discussed later in this paper.

We know that both pitchers and hitters have little effect on the batting average on balls in play (BABIP). Typically, this statistic will regress to around .300 over time. In plain English, a ball put in play results in a hit about 30% of the time, regardless of how well the ball is hit. Exceptions exist for both pitchers and hitters, but this information has shifted the focus of sabermetric analysis in baseball towards the analysis of outcomes that do not involve balls in play, and are therefore not subject to the inherent small-sample randomness of BABIP. These plays—strikeouts, walks, and homeruns—are known as the three true outcomes of baseball, and will be the basis for my analysis. I will be focusing on statistics that indicate the frequency at which a play ends in each of these outcomes—strikeout rate, walk rate, and home run rate—for both hitters and pitchers. Using these numbers, I will run regressions and attempt to establish the correlation between a player's statistics in foreign leagues and the NCAA to the player's statistics in the major leagues. I

believe there is a higher chance of strong correlation between the three true outcome statistics than other metrics.

I believe that the development of these sabermetrics and other statistics has already and will continue to expand the usefulness of statistical analysis in professional baseball, and could perhaps be the breakthrough needed to properly evaluate players in the leagues I am studying. As statistical analysis in sports continues to develop, there is a greater emphasis on metrics that identify specific player skills and measure them independent of the team environment around them. For example, in baseball, a statistic like runs batted in (RBI) would not be a good identifier of true player performance because his teammates hitting ahead of him must reach base and run the bases well for that player to be attributed an RBI. Batting average would be a bad statistic to study for my model because it takes too long to stabilize, and is highly correlated with BABIP. Strikeout rate, walk rate, and home run rate are measures of three specific skills for pitchers and hitters.

2. Literature Review

In 1996, Clay Davenport, co-founder of the sabermetric baseball website Baseball Prospectus, developed a context-independent metric called Equivalent Average (EqA), which attempted to put all baseball players in all leagues onto a level playing field. Ballparks, leagues, and platoon matchups can all be advantageous for a pitcher or a hitter, distorting statistics that do not account for these effects. A left-handed hitter may be used exclusively in platoon matchups against right-handed pitchers, which provides a substantial advantage for the hitter. A player might play in a ballpark like Coors Field, built in the high atmosphere of Denver, Colorado where the ball carries more, providing

another advantage for hitters. A ballpark can also be built to be pitcher-friendly, with long distances to high-standing walls providing for difficult home run targets, like San Francisco's AT&T Park. In addition to this, some leagues such as the Korean Baseball Organization (KBO) have reputations as hitter-friendly. EqA is an attempt to neutralize all of this and develop one metric for comparing baseball players should they all compete in the same environment against the same competition. Although scaled like batting average, EqA attempts to capture runs produced per at bat, something more along the lines of modern-day all-encompassing sabermetrics like weighted runs created (wRC) or weighted on-base average (wOBA). Although most commonly cited to compare Major League players from different eras, EqA can be used to measure how a foreign or minor league player might perform in a major league environment. Davenport's work will be the foundation of my study.

The specific metrics studied in this paper are worthy talking points. Franks, D'Amour, Cervone, and Bornn (2016) analyzed the effectiveness of various basketball and hockey statistics, using three meta-metrics by which they evaluate a statistics effectiveness. This same kind of process will be applied to determine which baseball metrics are the best to study. Specifically, Franks, D'Amour, Cervone and Bornn evaluate metrics for their stability, or whether or not they are predictive in nature. This is to say whether someone's performance metric in 2016 would help predict that same performance metric in 2017. Baseball is subject to small-sample volatility. Even commonly cited hitting metrics like on-base plus slugging (OPS) can be extremely volatile season-to-season. Adam Dunn was a productive hitter during his 14-year career. The worst offensive season of Dunn's career came in 2011, when he hit to a meager 54

OPS+ (OPS weighted with league average set equal to 100). Dunn never had another season, before or after 2011, with an OPS+ lower than 105. 2011 was the one and only season in which Dunn was a worse-than-league-average hitter, and in that one season, he was significantly worse than the league average. This speaks to not only the volatility inherent to the game of baseball, but also to the volatility in the sport's statistics. Of the three metrics I am focusing my model on, home run rate is subject to some season-to-season volatility, but strikeout rate and walk rate are extremely stable baseball metrics. I predict that, because of this, my model will more accurately predict performance based on strikeout rate and walk rate than home run rate.

One potential issue with my model is how performance can differ based on age. When comparing a player's age 22 season to his age 27 season, a bigger reason for difference in performance than the environment or the league he plays in could be the improvement a player has made in that time. For that reason, I will be attempting to incorporate age effects into my model. Fair (2008) estimated age effects for both hitters and pitchers in the same paper, but used on-base percentage (OBP), on-base plus slugging (OPS), and earned-run average (ERA). Much work on hitter and pitcher aging curves have been done independently of each other. Petti (2012) produced age curves for pitchers that use strikeouts per nine innings, walks per nine innings, and home runs per nine innings, very similar statistics to the ones I am studying. Zimmerman (2014) published aging curves for walk rate, strikeout rate and home runs per 600 plate appearances. Typically, peak performance for a hitter occurs between ages 26-28. However, specific hitter skills like walk rate and home run rate tend to increase until ages 30-32. Pitchers stay closer to peak performance a bit longer than hitters, but also

encounter a steeper decline as they age into their mid-to-late 30's. It is also important to keep the survivor bias in mind, which states that one way aging curves can be misrepresentative is because players who survive the big leagues in year one may have only survived because of small-sample luck, and as a result may see a false drop in performance in year two. Because of the issues and complications in implementing aging curves, I will simply be controlling for age in my regression equation.

The economic relevance of my model is based on proper valuation of what a free agent should be worth in professional baseball. Much work has been done on developing a model for this based on dollars spent on a free agent contract and a player's Wins Above Replacement (WAR), an all-encompassing metric that attempts to estimate a player's value in terms of wins based on his all-around performance. Although I am not studying WAR, some of the literature done on this topic will provide good context for the current state of player valuation in Major League Baseball. WAR is simply an estimation of overall player productivity, and there are multiple calculations of the metric. For the sake of the rest of my study, whenever referring to WAR, I will be using the Fangraphs calculation for WAR known as fWAR. Weinberg (2016) details a few key principles in how to evaluate a free agent contract. One is that teams sign players for future performance, not past performance. In the context of an international free agent, a team signs a player not for what he did in the KBO or the NPB, but for what a team thinks that player can do at the Major League level. This is the relevance of my model; teams are not going to sign players based on how good their numbers were in another league, teams are going to sign players based on how they think those numbers will translate to the major leagues. That is where a predictive model like mine can come in handy. Another

important factor is inflation, which will need to be accounted for when evaluating the efficiency with which teams have handed out contracts to international free agents in the past. A player making $8 million in 2006 is not the same as a player making $8 million in 2016, not just because of the inflation of the United States Dollar, but also due to the inflation of the labor market in professional baseball. The final key principle from Weinberg (2016) is that teams pay for an entire contract. A team may sign a player to a five-year deal worth $50 million, get a bargain in the first two years, and a fair level of performance in the following two before the contract becomes an albatross in the final year. At the end of the contract, if the total WAR of a player is worth more than the WAR that $50 million typically buys on the open market, no matter the distribution of performance, the contract was a good contract for the team. Swartz (2017) concluded that in 2017 teams spent on average $10.5 million per win in terms of WAR.

If a win is worth $10.5 million on the major league free agent market, then how have recent international free agents lived up to their financial billing? The international market for major leaguers has been extremely hit or miss. I have already discussed the example of Rusney Castillo, a major Cuban flop on the part of the Boston Red Sox. However, extremes exist on the other end of the spectrum as well. Jung-ho Kang signed with the Pittsburgh Pirates out of the KBO ahead of the 2015 season. Kang was one of the first Korean ballplayers to come up in the KBO and transition to the MLB, so the market was rightfully skeptical on Kang, and he received only a 4 year/$11 million contract. Despite missing the entire 2017 season for reasons other than performance, Kang amassed 6.0 fWAR in his first two Major League seasons, playing only 229 out of a possible 324 games. Six wins over the lifespan of a contract is worth $63 million based

on the model made by Swartz (2017). Kang accumulated that in the first two years of a deal worth just $11 million total. There is further evidence of the international free agent market being an inefficient one. Yoenis Cespedes signed for $36 million over four years before the 2012 season and tallied 15.3 fWAR over the life of his contract. Jose Abreu will enter the fifth year of his six-year contract in 2018, but has already collected 14.5 fWAR in his time at the major league level, a figure that would be worth $152.25 million on the open market today according to Swartz's model. There are also more examples of busts like Castillo, such as Daisuke Matsuzaka, who the Red Sox paid a total of $103 million for in both negotiating rights and contract fees before the 2007 season, and who totaled just 7.5 fWAR over the life his six-year contract. Based on this, one can conclude major league teams have had difficulty in evaluating the international free agent market. The goal of my model is to make this market a more efficient one.

Russell Carleton originally published a piece on stabilization points for some baseball metrics in 2007 on StatSpeak.net. His article was re-published on Fangraphs in 2011 and was further summarized in the website's library. Carleton (2007) discusses the issues with evaluating baseball players based on statistics measured in small sample sizes, and attempts to establish a cutoff point where these statistics are able to quantify true performance. It should be noted that these stabilization points are not magic numbers where these metrics all of a sudden become stable. Rather, these metrics will get more and more stable as the sample of plate appearances/batters faced grows, and the stabilization points are simply baselines. He uses the threshold of where R-squared = 0.49, or where the correlation coefficient of one sample of plate appearances and another sample of the same size equals 0.7 and where greater than 50% of the variance within the

sample is stable, as the point of stabilization for baseball statistics. Carleton also reveals that different metrics will have different stabilization points. For batters, strikeout rate stabilizes at 60 plate appearances, walk rate at 120 plate appearances, and home run rate at 170 plate appearances. For pitchers, strikeout rate stabilizes at 70 batters faced, walk rate at 170 batters faced, and home run rate at 1,320 batters faced. Carleton encounters a similar issue to one that my study has: anytime a minimum sample requirement is used, it becomes a selective sample. Playing time is not distributed randomly; better players play more, and as a result, become a bigger part of the sample.

3. Data

The foreign and amateur leagues I analyze are the Korean Baseball Organization (KBO), Nippon Professional Baseball (NPB), and the National Collegiate Athletic Association (NCAA). The NPB and KBO are two of the top foreign leagues from which MLB teams will consider signing professionals. Teams draft players out of the NCAA, which will be the only amateur league I study. These represent some of the major player pools from which major league teams consider players for their organizations. Players are also drafted out of high school, but statistics from high school leagues are difficult to come across, and attempting to quantify the wide range of competition across high school baseball in the United States would be extremely onerous.

The metrics studied are strikeout rate, walk rate, and home run rate. These encapsulate baseball's three true outcomes, which are the metrics hitters and pitchers have the most control over, and therefore are the most important to study when evaluating player performance.

I have gathered data from a few different sources for use in this project. I am using the Sean Lahman data set for all of my MLB data. The Lahman data provides statistics throughout all of major league history through the 2016 season. I gathered KBO and NPB data from http://japanesebaseball.com/data/index.gsp. I am using the Batting, Pitching, and Players sheets from the Pro Yakyu Database by Michael Westbay for the NPB, and the Master, Batting, and Pitching data sheets from user Beemer's Korean Baseball Database. Both my KBO and NPB data only go through the 2008 season, so I added any players who switched to the MLB after that using tables from baseballreference.com. For NCAA data, I purchased both batting and pitching data for the 2013-2017 seasons from http://www.thebaseballcube.com/. This means my NCAA data will consist of all batters and pitchers who played an NCAA season in 2013 or later and have since debuted in the MLB.

Each data source came with sets of batter statistics, pitching statistics, and a master set with all players and matching player identifications. For each league being studied, I joined the batting and pitching data with the master data to match players with their statistics. I then combined the plate appearances/batters faced, strikeouts, walks, and home runs for each player for the first three MLB seasons and final three non-MLB seasons from all leagues being studied, and calculated strikeout rate, walk rate, and home run rate for each player to use as variables in my regressions.

Only the NCAA data came with an age variable, so I manually estimated age for much of this data. I did so by subtracting the season year from a player's birth year.

I have also limited the data to only the three seasons before and after a player made the transition from a foreign or amateur league to the MLB. For example, Akinori

Iwamura left the NPB for the MLB after the 2006 season. This means his 2004-2006 seasons in the NBP and his 2007-2009 seasons in the MLB are the ones accounted for in my data. I do this in order to account for the natural progression or regression of a player's skill set over time. Constraining the timeframe to within a set window centered on a player's last year before transitioning to the MLB paints a clearer statistical picture for how much a player's performance is affected.

The criteria for a player to be included in the data is 170 engagements—plate appearances for hitters and batters faced for pitchers—over the course of the three-year window in one of the non-MLB leagues being studied, followed by at least 170 engagements in the MLB over the following three-year window. The 170 figure is based on work done by Carleton (2007) detailed above in the literature review section. Carleton's initial figures published on StatSpeak.net in 2007 are slightly different from the ones Fangraphs' posted in its library in 2011. For the sake of my study, I will be using Fangraphs' updated versions of Carleton's stabilization points. According to Carleton (2007), the stabilization for a pitcher's home run rate is 1,320 batters faced, which poses an issue as no pitcher faces that many hitters over the course of a single season anymore, and even applying that criteria over a three-year window would severely limit the number of observations. Since the 170 threshold covers five of the six metrics I am studying, I am going ahead with 170 plate appearances/batters faced as my sample minimum.

Before applying Carleton's 170 rule, my data consisted of 140 batters. With the 170 rule, my batting data consists of 39 players (five from the KBO, seven from the NCAA, and 27 from the NPB.

KBO	NCAA	NPB
5	7	27

Statistic	N	Mean	St. Dev.	Min	Median	Max
Age	37	28.732	4.937	21.000	29.000	41.000
nonMLB_KRate	38	16.640	7.687	3.875	15.768	32.417
nonMLB_BBRate	38	13.893	4.398	4.720	14.368	23.206
nonMLB_HRRate	38	5.022	3.291	0.354	4.669	14.647
MLB_KRate	38	18.668	5.931	8.197	18.656	32.787
MLB_BBRate	38	7.960	2.887	3.689	7.999	14.673
MLB_HRRate	38	2.594	1.537	0.000	2.612	5.755

Table 1

The average non-MLB strikeout rate among these 38 players is 16.64%, while the average MLB strikeout rate is 18.67%. The maximum non-MLB strikeout rate is Darnell Coles at 32.41% and the minimum MLB strikeout rate is Tony Batista at 3.88%, while the maximum MLB strikeout rate is Byung Ho Park (32.79%) and the minimum MLB strikeout rate is Kenji Johima (8.2%). For walk rate, the non-MLB average is 13.89%, while the MLB average is 7.96%. The maximum non-MLB walk rate is Kosuke Fukudome at 23.21% and minimum the non-MLB walk rate is Hiroki Kuroda at 2.91%, while the maximum MLB walk rate is Fukudome (14.67%) and the minimum MLB walk rate is Johima (3.69%). Home run rate follows a similar trend to walk rate, as the average non-MLB home run rate is 5.02%, while the average MLB home run rate is 2.59%. The maximum non-MLB home run rate is Jolbert Cabrera's 14.65% and the minimum non-MLB home run rate is Kuroda's 0.34%, while the maximum MLB home run rate is Kyle Schwarber (5.76%) and minimum MLB home run rate is Kuroda and Tsuyoshi Nishioka (0%). None of these results are unexpected; assuming MLB is a more difficult league than the KBO, NPB and NCAA, and player performance should dip as a result, MLB strikeout rate should be higher than non-MLB strikeout rate, while MLB home run rate

and walk rate should be lower than non-MLB home run rate and walk rate. I see these trends play out in both the means and medians. The standard deviation on strikeout rate for both MLB and non-MLB data is the highest, while standard deviation on home run rate for MLB and non-MLB is the lowest. This makes sense, because strikeouts are the highest-frequency event of the three studied, while home runs are the lowest.

Without Carleton's 170 rule, my pitching data consisted of 149 pitchers. With it, my pitching data consists of 83 players (16 from the KBO, 18 from the NCAA and 49 from the NPB). It is interesting to note that this is more than twice the number of observations in this data relative to my batting data. It appears teams have been more willing to sign pitchers from foreign leagues, and more willing to rush college pitchers through the ranks of minor league baseball after being drafted than college batters.

KBO	NCAA	NPB
16	18	49

Statistic	N	Mean	St. Dev.	Min	Median	Max
Age	83	28.315	4.550	20.000	29.000	36.000
nonMLB_KRate	83	20.503	6.196	8.021	20.528	40.190
nonMLB_BBRate	83	7.253	2.543	3.097	6.700	14.573
nonMLB_HRRate	83	1.992	1.167	0.000	1.974	8.411
MLB_KRate	83	17.792	4.700	9.610	16.842	32.907
MLB_BBRate	83	8.393	2.392	3.736	8.271	14.545
MLB_HRRate	83	3.142	1.052	1.128	2.946	6.061

Table 2

The average non-MLB strikeout rate in this data is 20.5%, while the average MLB strikeout rate is 17.79%. The maximum non-MLB strikeout rate is 40.19% (Kazhiro Sasaki) and minimum non-MLB strikeout rate is 8.021% (Mike Fyhrie), while the

maximum MLB strikeout rate is 32.91% Seung-Hwan Oh and minimum MLB strikeout rate is 9.16% (Brian Sweeney). The average non-MLB walk rate is 7.25%, while the average MLB walk rate is 8.39%. The maximum non-MLB walk rate is Wes Obermueller's 14.573% and minimum non-MLB walk rate is Kyle Crockett's 3.1%, while the maximum MLB walk rate is Kuzhisa Ishii's 14.55% and minimum MLB walk rate is Koji Uehara's 3.74%. The average non-MLB home run rate is 1.99%, while the average MLB home run rate is 3.14%. The maximum non-MLB home run rate is 8.41% (Pete Walker) and minimum non-MLB home run rate is 0% (Crockett and Marco Gonzales), while the maximum MLB home run rate is 6.06% (Chad Green) and minimum MLB home run rate is 1.13% (Crockett). These results are a bit less expected than the results of the batting data. For one, the walk rate is very similar for MLB and non-MLB data in terms of standard deviation, maximum and minimum, while the means are also close. The maximum non-MLB home run rate is higher than the maximum MLB home run rate, which is also unexpected, but the median and mean is lower in the non-MLB group, which is expected. The standard deviations follow the same pattern as the batting data. The means follow the inverse patterns of the batting data, as expected, where strikeout rate is lower in the MLB, while walk rate and home run rate is higher in the MLB.

4. Model Design

I will be running regression models to determine the correlation between performance metrics in the top foreign/amateur leagues and Major League Baseball. The end goal for my model will be to predict a player's strikeout rate, walk rate and home run rate in the MLB based on his performance in terms of those same metrics in another

league. I may find that certain metrics are undervalued when it comes to statistical analysis performed on players from outside leagues, or that statistical analysis is indeed largely useless when it comes to evaluating players across environments. Not only will my model aim to help better predict the performance of high-priced international free agents and top college draft picks, but it should also help uncover hidden gems, or players who are being wrongly overlooked by professional teams despite strong underlying metrics.

I limit my inputs to exclusively players who made the transition from a foreign, minor, or amateur league to the major leagues, and not the other way. It is common for a player at the tail end of his career to play overseas in hopes of extending his career. However, that data will not be applicable here, as I am developing a model for predicting performance in the MLB based on statistics from other leagues, not the other way around. I also limit my output MLB data to the first three years of a player's major league career. This is because, at the end of year three, it is reasonable to expect other major league evaluators to evaluate a foreign player exclusively on his major league numbers.

5. Model

I run regressions for each of the three metrics detailed above. The dependent variable is the metric for a hitter or pitcher in a major league environment, while the independent variables are the same metric for a hitter or pitcher in another environment, as well as a combination of league dummy variables, the interactions between the dummies and the non-MLB metric, and age. The regression equation will look as such:

$$y_i = \beta_0 + \beta_1 * x_i + \beta_2 * NCAA + \beta_3 * NPB + \beta_4 * x_i * NPB + \beta_5 * x_i * NCAA + \beta_6 * Age$$

y_i represents performance for player i in Major League Baseball, while x_i represents performance through the same metric in another league. There is a set of dummy variables, with the complete set representing the NPB and NCAA, while KBO is the omitted dummy. For each regression, the hypotheses will be as follows:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

The null hypothesis is that the coefficient on the non-MLB performance metric is zero, indicating no relationship between Major League performance and performance in other leagues. I use a two-sided alternative hypothesis because the relationships between the metrics could be positive or negative. β_1 is my correlation coefficient that can be used to forecast statistical translations after the model is developed. B_2 and β_3 are dummy variables that differentiate which league I am comparing to the MLB, with all leagues except KBO garnering a dummy variable coefficient. When all dummy variable coefficients are zero, KBO is the league being measured against the MLB. β_4 and β_5 measure the interactions between the league dummies and the studied metrics, allowing me to differentiate the correlation coefficients for different leagues. β_6 is an attempt to control for the natural aging curve that occurs in athletes through an age variable. We can expect that a player's performance in professional baseball will peak in his late 20's. If a player played in a foreign league in his 20's and the major leagues in his 30's, his performance is expected to dip even further than usual. I estimate separate models for pitchers and hitters and for each different metric. I only evaluate players through the 2016 season, since that is where the Lahman data ends at the time of this project.

6. Model Results

	Dependent variable:					
	MLB_KRate (1)	MLB_BBRate (2)	MLB_HRRate (3)	MLB_KRate (4)	MLB_BBRate (5)	MLB_HRRate (6)
nonMLB_KRate	1.436***			0.566***		
	(0.479)			(0.156)		
nonMLB_BBRate		0.281			0.673***	
		(0.355)			(0.184)	
nonMLB_HRRate			0.627***			0.354
			(0.219)			(0.439)
dummyNCAA	21.742*	-1.660	3.584**	7.400	3.665	0.608
	(11.457)	(5.348)	(1.486)	(5.287)	(2.562)	(1.019)
dummyNPB	19.676**	-1.815	1.622	2.978	2.102	-0.060
	(8.245)	(4.703)	(1.182)	(3.579)	(1.777)	(0.947)
nonMLB_KRate:dummyNCAA	-1.035			-0.319		
	(0.873)			(0.234)		
nonMLB_KRate:dummyNPB	-1.449***			-0.086		
	(0.494)			(0.181)		
nonMLB_BBRate:dummyNCAA		0.064			-0.383	
		(0.418)			(0.352)	
nonMLB_BBRate:dummyNPB		0.007			-0.249	
		(0.377)			(0.217)	
nonMLB_HRRate:dummyNCAA			-0.490*			-0.268
			(0.285)			(0.686)
nonMLB_HRRate:dummyNPB			-0.518**			-0.210
			(0.229)			(0.457)
Constant	-2.437	5.369	-0.049	5.527*	3.094**	2.693***
	(7.884)	(4.299)	(1.104)	(3.031)	(1.538)	(0.881)
Observations	38	38	38	83	83	83
R2	0.382	0.218	0.462	0.382	0.270	0.057
Adjusted R2	0.286	0.096	0.378	0.342	0.222	-0.004
Residual Std. Error	5.012 (df = 32)	2.746 (df = 32)	1.213 (df = 32)	3.814 (df = 77)	2.109 (df = 77)	1.054 (df = 77)
F Statistic	3.964*** (df = 5; 32)	1.782 (df = 5; 32)	5.488*** (df = 5; 32)	9.506*** (df = 5; 77)	5.684*** (df = 5; 77)	0.936 (df = 5; 77)
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 3

Table 3 shows the results of my regressions. There are six specifications, one for each of the three dependent variables (MLB strikeout rate, MLB walk rate, MLB home run rate) with both the batters and pitchers data. The independent variables are a combination of the same metrics in the non-MLB leagues, dummies, and dummy interactions. I dropped the age variable from my models because its inclusion lowered the adjusted R-squared values in each of my regressions with the batters data. The table shows that there is a positive and statistically significant relationship between non-MLB performance and MLB performance in most cases. The coefficient on non-MLB strikeout rate for batters in the first regression is statistically significant and equal to 1.4, so a one-

percentage point increase in a player's non-MLB strikeout rate correlates to a 1.4 percentage point increase in the MLB, holding all other variables constant. The regression also includes the interactions of non-MLB strikeout rate with the NCAA and NPB dummies. Therefore, the interpretation of the 1.4 coefficient is the effect of non-MLB performance on MLB performance for KBO players. The coefficient on the interaction between strikeout rate and the NCAA dummy is statistically insignificant, indicating that NCAA strikeout rate has the same predictive power as KBO strikeout rate. The coefficient on the interaction between strikeout rate and the NPB dummy is -1.4, offsetting the coefficient on non-MLB strikeout rate, suggesting that NPB strikeout rate does not predict MLB strikeout rate. The non-MLB walk rate coefficient for batters is not statistically significant, nor are the interaction terms significant, indicating that there is no relationship between the walk rates in the foreign/amateur leagues and MLB walk rates. For non-MLB home run rate, the coefficient is statistically significant and equal to 0.6, meaning a one-percentage point increase in KBO home run rate corresponds to a 0.6 percentage point increase in MLB home run rate. The variable representing interaction between the NCAA and non-MLB home run rate is also statistically significant and equal to -0.5, so a one-percentage point increase in NCAA home run rate correlates with a 0.1 percentage point increase in MLB home run rate. Likewise, the NPB and non-MLB home run rate interaction variable is statistically significant and equal to -0.5, meaning a one percentage point increase in NPB home run rate correlates to a 0.1 percentage point increase in MLB home run rate.

In the pitchers data, the coefficient on non-MLB strikeout rate is statistically significant and equal to 0.6. Holding all other variables constant, a one-percentage point

increase in player's KBO strikeout rate is correlates with a 0.6 percentage point increase in his MLB strikeout rate. The interactions between the NPB/NCAA dummies and non-MLB strikeout rate for pitchers are not statistically significant, so we can conclude a one-percentage point increase in strikeout rate in these leagues also corresponds to a 0.6 percentage point increase in MLB strikeout rate. The constant in this regression is statistically significant and equal to 5.5, so a pitcher's strikeout rate is expected to increase 5.5 percentage points if non-MLB strikeout rate is zero. The coefficient on non-MLB walk rate is 0.7 and statistically significant, so a one-percentage point in KBO walk rate correlates with a 0.7 increase in MLB walk rate. Again, the interactions between the NPB/NCAA dummies and non-MLB walk rate are not statistically significant, so I assume a one-percentage point increase in walk rate in these leagues also corresponds with a 0.7 increase in MLB walk rate. The constant is statistically significant and equal to 3.1, so a pitcher's walk rate is expected to increase by 3.1 percentage points should the non-MLB walk rate be zero. The coefficient on non-MLB home run rate is not statistically significant, so it does not appear that non-MLB home run rate is predictive of MLB home run rate for pitchers. The constant on this model is statistically significant, equaling 2.7, so a pitcher's home run rate should increase by 2.7 percentage points if non-MLB home run rate is zero.

The adjusted R-squared value on my first estimation is 0.286, meaning my regression explains only 28.6% of the variation in MLB strikeout rate. Figure 1 shows the relationship between non-MLB strikeout rate and MLB strikeout rate among batters visually, and it does not appear to be strong.

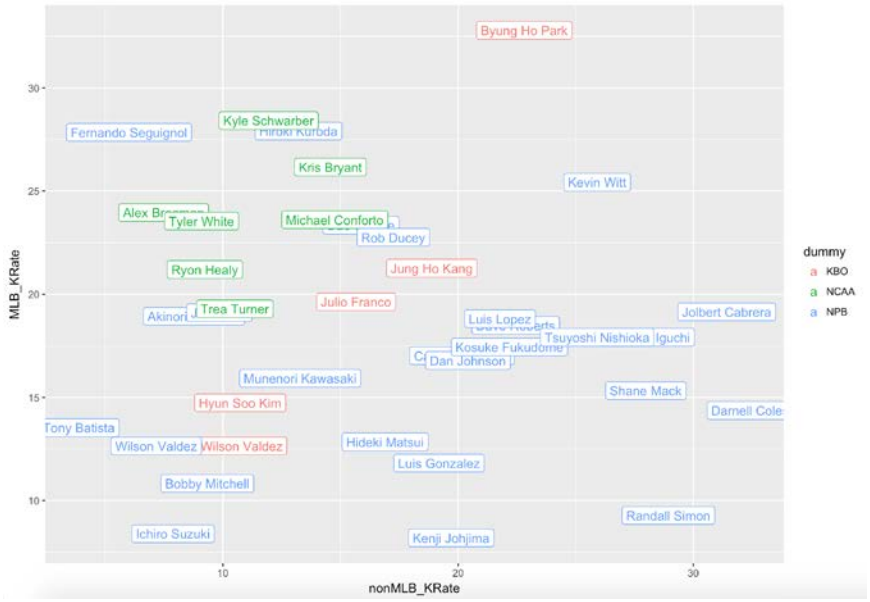

Figure 1

Figure 1 illustrates a stronger relationship in non-MLB strikeout rate and MLB strikeout rate for the KBO and NCAA, which fits in line with the observations from Table 3. In the walk rate estimation for batters, the adjusted R-squared is 0.096, meaning this estimation explains only 9.6% of the variation in MLB walk rate. Looking at a graph of the walk rate data, it does not appear that the walk rate relationship is particularly weaker in any of the leagues, rather noisy throughout the entire data.

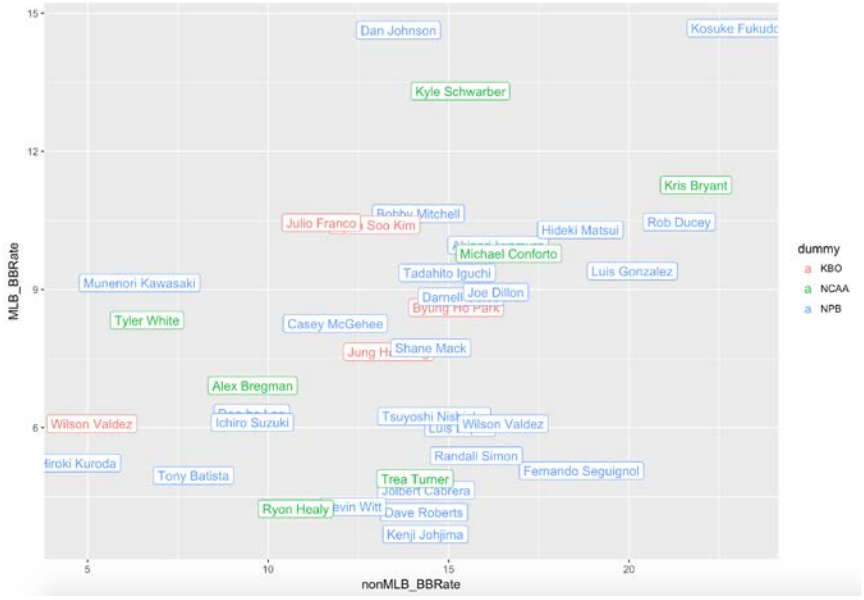

Figure 2

The adjusted R-squared for the home run rate model in the batters data is 0.378, meaning this regression accounts for 37.8% of the variation in MLB home run rate. Analyzing the graph details further insight I can garner from this data.

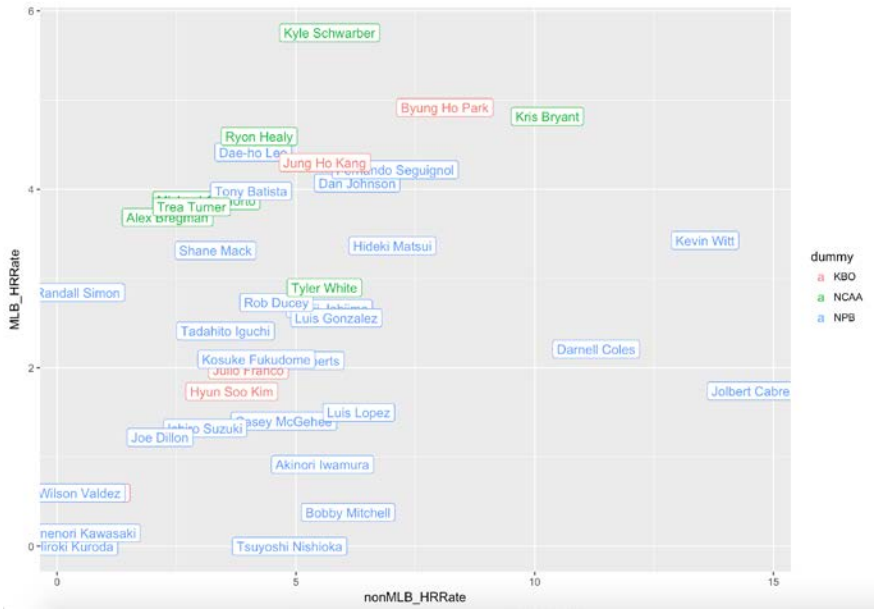

Figure 3

The data is relatively compact along the x-axis aside from three outliers that fall significantly to the right of the rest of the data. Darnell Coles, Kevin Witt, and Jolbert Cabrera each have high NPB home run rates (all above 10%) and low MLB home run rates (all below 4%). Perhaps these outliers are due to small sample size. Each of the three outlier players from the NPB accumulated more than 500 plate appearances in the NPB over the three-year window being studied, but Coles and Witt fell short of 400 plate appearances in the MLB over the following three-year window, and Cabrera only amassed 517 MLB plate appearances. Each of these three players reached the 170 plate appearances stabilization point for a hitter's home run rate outlined in Carleton (2007).

Despite these outliers, I am still able to develop a model with non-MLB home run rate as a statistically significant predictor of MLB home run rate.

Looking again at the pitchers data, the adjusted R-squared value for the model analyzing strikeout rate is 0.342, this model can account for meaning 34.2% of the variation in MLB strikeout rate. The graph of non-MLB strikeout rate and strikeout rate for pitchers indicates a fairly linear relationship between the metrics.

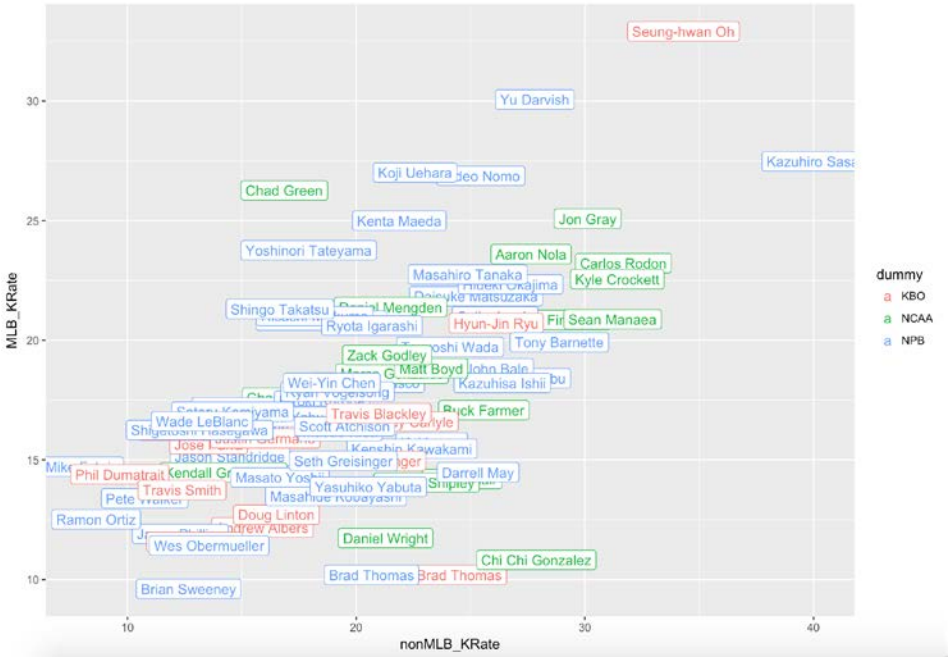

Figure 4

The non-MLB and MLB relationship for this metric appears particularly strong based on the graph. This played out in the model results with a lower p-value for non-MLB strikeout rate for pitchers than seen in the other models. Unlike when regressing non-MLB strikeout rate on MLB strikeout rate for batters, there is not one league that stands out as having a weaker relationship than the others as observed with the NPB in the batters data. The adjusted R-squared on the walk rate model for pitchers is relatively low at just 0.222 meaning this regression accounts for only 22.2% of the variation in

MLB walk. Figure 5, the graph of non-MLB walk rate and MLB walk rate from the pitchers data, is not as tight around the center as the graph in Figure 4. This graph gets less compact as the non-MLB walk rate increases. One potential conclusion to draw from this is that non-MLB walk rate is only particularly useful as a predictor when it is below a certain threshold.

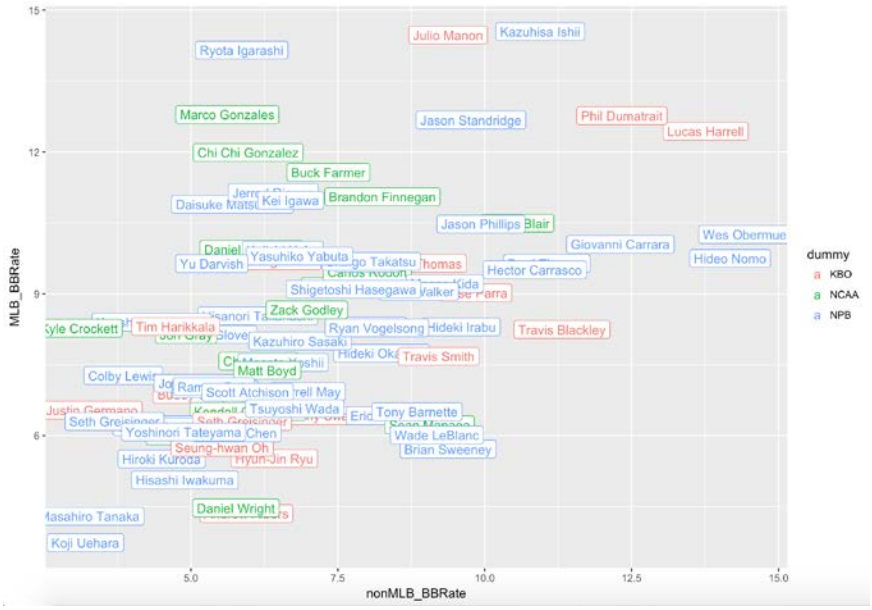

Figure 5

The adjusted R-squared value for the home run rate model is -0.004, and my model does not account for any of the variation in MLB home run rate. This makes sense since the variable non-MLB home run rate was not statistically significant.

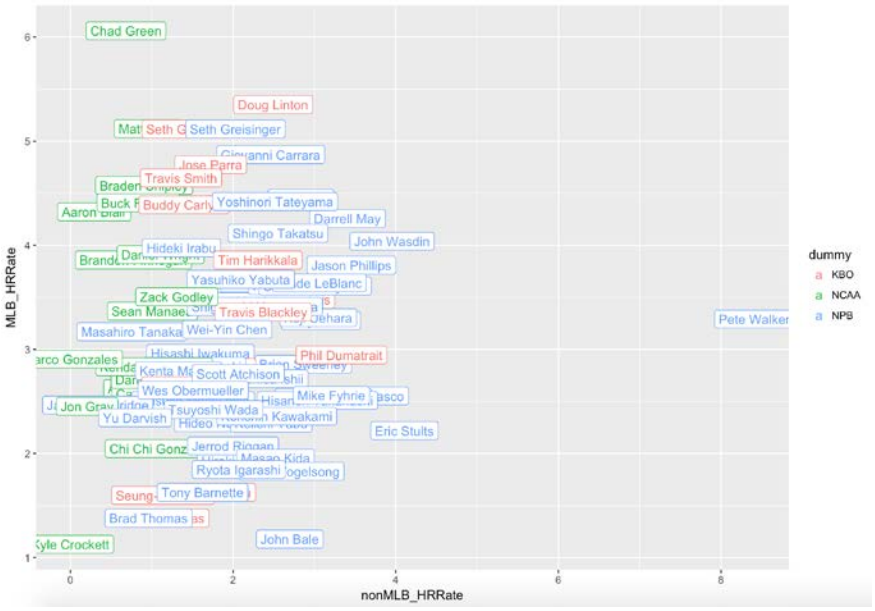

Figure 6

Figure 6 shows the plot of non-MLB and MLB home run rates for pitchers and details a sharp upward-sloping relationship between these metrics aside from one major outlier, which lies significantly rightward of the group. Walker's metrics are drawn from a sample of 1,337 MLB batters faced but only 214 non-MLB batters faced. Walker's inclusion makes the best fit horizontal rather than vertical. This is an example of the issue with using the 170 rule on a pitcher's home run rate, as it appears Walker's home run rate in the NPB failed to stabilize, sitting extremely high at 8.41%, whereas his MLB home run rate settled at a much more reasonable 3.29%.

7. Using Playing Time as a Performance Metric

I also decided to do some estimations on MLB plate appearances and MLB batters faced. These measures of engagements can represent a performance metric. In theory, the best players are going to get more chances to play, and will accumulate more plate appearances/batters faced over time as a result. There are, however, a number of factors go into playing time, not just performance and ability. A player might find himself

with more opportunities on an MLB team with inferior players at the position he plays simply because of the situation he is placed in. The data I am working with does not cover all of these factors, but the analysis is still worthwhile. I am able to ignore the 170 rule in this analysis, and as a result, my batters data increases to 140 players, while my pitchers data increases to 149 players. For this analysis, it is important to keep in mind that I am no longer looking at rates, rather the raw volume of plate appearances/batters faced. It should be noted that the season lengths for these leagues are all different; MLB plays a 162-game regular season, whereas the NPB is a 143-game regular season, the KBO is a 144-game regular season, and the NCAA is about a 60-game regular season, varying from team to team. Table 4 shows the results of three regressions run on the batters data.

	Dependent variable:		
	mlb_PA		
	(1)	(2)	(3)
PA	0.358***	0.397***	0.361
	(0.071)	(0.079)	(0.309)
nonMLB_KRate		0.003	4.550
		(0.024)	(30.551)
nonMLB_BBRate		0.021	-27.936
		(0.032)	(37.961)
nonMLB_HRRate		0.018	-53.858
		(0.071)	(103.777)
dummyNCAA		-137.129	-993.278*
		(211.873)	(588.196)
dummyNPB		71.577	-430.905
		(143.543)	(418.763)
Age		-10.053	-12.707
		(12.480)	(12.818)
PA:dummyNCAA			-0.107
			(0.656)
PA:dummyNPB			0.080
			(0.322)
nonMLB_KRate:dummyNCAA			-1.344
			(31.289)
nonMLB_KRate:dummyNPB			-4.545
			(30.551)
nonMLB_BBRate:dummyNCAA			44.865
			(50.182)
nonMLB_BBRate:dummyNPB			27.959
			(37.961)
nonMLB_HRRate:dummyNCAA			112.822
			(116.048)
nonMLB_HRRate:dummyNPB			53.877
			(103.778)
Constant	141.128**	358.714	912.271
	(55.308)	(410.061)	(578.909)
Observations	140	139	139
R2	0.154	0.188	0.230
Adjusted R2	0.148	0.144	0.136
Residual Std. Error	490.788 (df = 138)	493.366 (df = 131)	495.694 (df = 123)
F Statistic	25.076*** (df = 1; 138)	4.327*** (df = 7; 131)	2.452*** (df = 15; 123)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4

The first model is a regression on MLB plate appearances with plate appearances in the foreign or amateur league as the only independent variable. There is a statistically significant and positive relationship between plate appearances in the non-MLB leagues

and MLB. The coefficient on the independent variable is 0.4, meaning an increase in non-MLB plate appearances by one correlates with a 0.4 unit increase in MLB plate appearances. The constant is also statistically significant and equal to 141.1, so MLB plate appearances are expected to increase by 141.1 if non-MLB plate appearances is zero. The adjusted R-squared on this model is 0.148, meaning non-MLB playing time explains about 14.8% of the variance in MLB playing time. My second estimation does not include the interaction variables, and the foreign/amateur plate appearance variable is statistically significant with a positive relationship again. The coefficient is equal to 0.4, so increasing non-MLB plate appearances by one corresponds to a 0.4 increase in MLB PAs. The adjusted R-squared of this model is 0.144, so this estimation accounts for 14.4% of the variance in MLB plate appearances. This is not as high as the original model with just foreign/amateur plate appearances as the lone independent variable. My final estimation includes the foreign/amateur plate appearances as well as the foreign/amateur performance metrics, dummy variables, age variable, and the interactions between the performance metrics and dummies. I lose the statistical significance of foreign/amateur plate appearances and the adjusted R-squared dips to 0.136, so the model only accounts for 13.6% of the variance in my MLB plate appearances. Adding in these extra independent variables did not increase the model's usefulness, and the interactions did not help differentiate the translations between leagues. It does not appear that these performance metrics in foreign and amateur leagues are good predictors of playing time in the MLB for batters.

	Dependent variable:		
	mlb_BF		
	(1)	(2)	(3)
BF	0.461***	0.364***	0.308
	(0.067)	(0.073)	(0.304)
nonMLB_KRate		7.220	6.069
		(7.236)	(24.728)
nonMLB_BBRate		-9.627	21.697
		(13.162)	(44.574)
nonMLB_HRRate		49.224	71.728
		(30.797)	(226.995)
dummyNCAA		-637.880***	-293.242
		(200.392)	(970.292)
dummyNPB		137.955	328.215
		(132.956)	(909.119)
Age		-54.961***	-58.803***
		(15.210)	(16.524)
BF:dummyNCAA			0.085
			(0.450)
BF:dummyNPB			0.060
			(0.314)
nonMLB_KRate:dummyNCAA			0.590
			(27.528)
nonMLB_KRate:dummyNPB			2.078
			(26.907)
nonMLB_BBRate:dummyNCAA			-41.656
			(48.872)
nonMLB_BBRate:dummyNPB			-29.722
			(48.198)
nonMLB_HRRate:dummyNCAA			-101.422
			(268.003)
nonMLB_HRRate:dummyNPB			-17.140
			(229.762)
Constant	251.337***	1,783.848***	1,662.650
	(63.383)	(544.413)	(1,018.418)
Observations	149	148	148
R2	0.241	0.352	0.358
Adjusted R2	0.236	0.320	0.285
Residual Std. Error	532.383 (df = 147)	502.913 (df = 140)	515.536 (df = 132)
F Statistic	46.681*** (df = 1; 147)	10.872*** (df = 7; 140)	4.910*** (df = 15; 132)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5

In my first estimation from Table 5, regressing MLB batters faced on non-MLB batters faced, I see similar results to those from my batters data. There is a statistically

significant and positive relationship between foreign/amateur batters faced and MLB batters faced. The coefficient is equal to 0.5, so increasing foreign/amateur batters faced by one unit corresponds to a 0.5 unit increase in MLB batters faced. The constant is statistically significant and equal to 251.4, so a pitcher's batters faced is expected to increase by an average of 251.4 MLB should the non-MLB batters faced total equal zero. It is interesting that the adjusted R-squared for this model is significantly higher than the first model estimating MLB plate appearances with the batters data. In this case, it is 0.236, meaning non-MLB playing time accounts for about 23.6% of the variation in MLB playing time among pitchers. The second model gives me statistically significant non-MLB batters faced, age, dummyNCAA variables and constant. The coefficient on non-MLB batters faced is 0.4, so increasing non-MLB batters faced by one correlates with a 0.4 increase in MLB BFs. The constant coefficient jumps to 1,783.8, so a pitcher with zero non-MLB batters faced would an average of 1,783.8 MLB batters faced according to this estimation. The adjusted R-squared is up to 0.32, so this estimation accounts for 32% of the variation in MLB batters faced. I am observing a pattern where age appears to affect performance in pitchers more than in batters. In the final estimation, foreign/amateur batters faced is not statistically significant, but the age variable is, and the adjusted R-squared falls to 0.2852, meaning the model accounts for 28.52% of variation in MLB batters faced.

8.1 How effective is predicting future MLB performance with past MLB performance?

My models turned out to be relatively non-predictive of MLB performance for players making the transition from the NCAA, NPB and KBO. Limiting performance to a

six-year window, the strikeout rate, walk rate, and home run rate for batters and pitchers in the final three years in a foreign/amateur league do not strongly correlate to the same metrics for the first three years of a player's MLB career. This brings up the question of what actually does predict MLB performance. More specifically, is past performance a good predictor of future performance? I decided to do a similar regression analysis for future MLB performance on past MLB performance. I used the Sean Lahman data and calculated the strikeout rate, walk rate, and home run rate metrics using strikeouts, walks, home runs, and plate appearances. I applied the 170 rule based on the work done by Carleton (2007), and constrained my evaluation to a similar six-year window where I used the first three years of a player's MLB career to predict the next three years in terms of strikeout rate, walk rate, and home run rate.

	Dependent variable:		
	post_KRate	post_BBRate	post_HRRate
	(1)	(2)	(3)
pre_KRate	0.929***		
	(0.009)		
pre_BBRate		0.853***	
		(0.012)	
pre_HRRate			0.905***
			(0.011)
Constant	0.569***	1.747***	0.278***
	(0.121)	(0.099)	(0.023)
Observations	2,710	3,151	3,151
R2	0.811	0.619	0.698
Adjusted R2	0.811	0.619	0.698
Residual Std. Error	2.617 (df = 2708)	1.978 (df = 3149)	0.831 (df = 3149)
F Statistic	11,617.330*** (df = 1; 2708)	5,118.122*** (df = 1; 3149)	7,285.764*** (df = 1; 3149)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6

This is the result for batters. There are 3,151 observations for the walk rate and home run rate models, and 2,710 for the strikeout rate models. There were about 390 NAs for my pre_KRate variable, and 318 for the post_KRate variable, some of which

overlapped, explaining the difference in observations. This was due to some missing values in the Lahman data. I used the metrics for seasons 4-6 of a player's career as the dependent variables, and the metrics for seasons 1-3 of a player's career as the only independent variables, replicating the three-year window used in my foreign and amateur league models.

Each of my independent variables in the three estimations are statistically significant with p-values less than 0.01. The first model's coefficient is equal to 0.9, so a one percentage point increase in strikeout rate in a three-year period of a player's career corresponds to a 0.9 percentage point strikeout rate increase in an ensuing three-year period. The constant is statistically significant and equal to 0.6, suggesting that strikeout rate increases by 0.6 percentage points in an ensuing three-year period regardless of performance in the prior three-year period. The coefficient on the walk rate model is equal to 0.9, meaning a one percentage point increase in a player's walk rate in a three-year window correlates with a 0.9 percentage point increase walk rate for the ensuing three-year period. The constant in this model is statistically significant and equal to 1.7, indicating walk rate in an ensuing three-year period increases by 1.7 percentage points no matter performance in the prior three-year window. The coefficient on the home run rate model is also 0.9; a one percentage point increase in a player's home run rate in a three-year window corresponds with a 0.9 percentage point increase home run rate for the ensuing three-year period. The constant here is statistically significant and equal to 0.3, meaning home run rate should increase by 0.3 percentage points in an ensuing three-year window regardless of performance in the prior three-year window. Each of these

coefficients indicate nearly one-to-one relationships between these metrics from one three-year period to the next.

The adjusted R-squared value for regression of future strikeout rate on past strikeout rate is 0.811, meaning past strikeout rate can account for 81.1% of the variation in future strikeout rate. The adjusted R-squared value for doing the same with walk rate was also relatively high at 0.619, indicating past walk rate for batters accounts for 61.9% of the variation in future walk rate. For home run rate, the adjusted R-squared was 0.698; I can account for 69.8% of the variation in future home run rate with past home run rate.

These R-squared values, even in the home run rate model, are much higher than the ones I got when estimating with foreign and amateur stats. The adjusted R-squared on the foreign/amateur strikeout rate model for batters was 0.286, much lower than the 0.811 figure for the MLB on MLB model. On the walk rate model for batters, the adjusted R-squared was 0.096, again significantly lower than the 0.619 figure obtained in this section's walk rate model. The adjusted R-squared on the home run rate model for batters was 0.378, lower than the 0.698 number for this home run rate model. I can conclude that past MLB performance for batters is much more predictive of future MLB performance than past performance in foreign and amateur leagues.

	Dependent variable:		
	post_KRate	post_BBRate	post_HRRate
	(1)	(2)	(3)
pre_KRate	0.914***		
	(0.012)		
pre_BBRate		0.635***	
		(0.016)	
pre_HRRate			0.746***
			(0.016)
Constant	1.061***	2.910***	0.605***
	(0.182)	(0.145)	(0.032)
Observations	2,545	2,545	2,545
R2	0.694	0.387	0.467
Adjusted R2	0.694	0.387	0.466
Residual Std. Error (df = 2543)	3.174	1.794	0.704
F Statistic (df = 1; 2543)	5,768.132***	1,604.192***	2,224.007***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 7

Table 7 shows the results of similar estimations with my MLB pitchers data. This time, I have 2,545 observations for each regression. The coefficient on past strikeout rate is statistically significant and equal to 0.9, meaning a one percentage point increase in strikeout rate for a three-year period correlates to a 0.9 percentage point increase in strikeout rate in the ensuing three-year period. This is the only pitcher metric with a near one-to-one relationship between past performance and future performance. The constant is equal to 1.1, so future strikeout rate is expected to increase by 1.1 percentage points regardless of past strikeout rate. The coefficient on walk rate is statistically significant and equal to 0.6, so a one percentage point increase in past walk rate corresponds to a 0.6 percentage point increase in future walk rate. The constant is equal to 2.9; future walk rate is expected to increase by 2.9 percentage points disregarding past performance. The coefficient on past home run rate is statistically significant and equal to 0.7, indicating a

one percentage point increase in home run rate for one three-year period correlates to a 0.7 percentage point increase in home run rate for the ensuing three-year period. The constant equals 0.6, suggesting future home run rate will increase by 0.6 percentage points regardless of past performance.

The adjusted R-squared values are a bit lower this time; for regressing future strikeout rate on past strikeout rate, the adjusted R-squared is 0.694, meaning past strikeout rate accounts for 69.4% of the variation in future strikeout rate. The adjusted R-squared on the walk rate estimation is low at 0.387, so this estimation only accounts for 38.7% of the variation in future walk. On home run rate, the adjusted R-squared was 0.467; past home run rate accounts for 46.7% of the variation in future home run.

These adjusted R-squared figures are higher than the figures on the foreign/amateur models. The adjusted R-squared for the foreign/amateur strikeout rate pitchers model is 0.342, compared to 0.694 here. On the pitcher walk rate models, the foreign/amateur adjusted R-squared is 0.222, whereas on the MLB model it is 0.387. For the pitcher home run rate models, the foreign/amateur adjusted R-squared is zero, and here it is 0.467. I can conclude that pitcher performance, just like batter performance, is easier to predict with past MLB performance than with past performance in the KBO, NPB, or NCAA.

It appears that predicting future performance with past performance is more effective for batters than for pitchers, with a significant adjusted R-squared gap coming in walk rate, and another considerable adjusted R-squared gap in home run rate. Based on this analysis, I can pretty confidently predict future strikeout rate based on past strikeout rate, but at best predict only about 80% of the variation in future performance based on

past performance, and more realistically closer to 60%-70%. This helps put the low adjusted R-squared values in my estimations from Chapter 6 into context; past performance is not a perfect way to predict future performance.

8.2 Do existing prediction models outperform predictions based on past performance?

In this section, I compare an existing prediction model to my models in section 8.1, which is solely based on past performance. Prediction models like ZiPS, Steamer, and Depth Charts are published every year for predicting future MLB performance, but even these well-developed models are not perfect. Forecasting future performance is an inexact science; if it were exact, there would be no arbitrage, and teams would value every player the same as everyone else does.

Dan Szymborski, the creator of the ZiPS projection system, was kind enough to give me data with the ZiPS projections dating back to the 2015 season. I combined this with the Lahman MLB data, which is updated through 2016, to perform an analysis of how well Szymborski was able to predict performance in the 2015 and 2016 MLB seasons with his ZiPS projections. For the analysis, I used the following estimator for batters faced, which was not included in Szymborski's projections, and was needed to calculate the strikeout rate/walk rate/home run rate for pitchers. This specific estimator was suggested by Szymborski himself, and is endorsed by Voros McCracken, the creator of Defense Independent Pitching Statistics (DIPS) and a pioneer of baseball sabermetrics.
Batters faced=(((Innings Pitched*3)-Strikeouts)*.966)+Hits+Walks+Strikeouts

For some background on Szymborski's methodology, he uses weighted averages of four years of past performance (8/5/4/3), or three years for players at the extreme ends

of aging curves (very young or very old), and regresses past performance on DIPS theory and BABIP theory, while also incorporating age effects based on historical players with similar statistical profiles.

	Dependent variable:		
	KRate	BBRate	HRRate
	(1)	(2)	(3)
pred_KRate	0.892***		
	(0.023)		
pred_BBRate		0.877***	
		(0.031)	
pred_HRRate			0.913***
			(0.033)
Constant	2.624***	1.426***	0.598***
	(0.478)	(0.245)	(0.094)
Observations	714	714	714
R2	0.674	0.531	0.519
Adjusted R2	0.674	0.531	0.518
Residual Std. Error (df = 712)	3.337	2.074	1.108
F Statistic (df = 1; 712)	1,474.162***	806.743***	768.126***
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 8

Table 8 shows the results of regressing strikeout rate, walk rate, and home run rate for batters on the ZiPS projections from the 2015 and 2016 seasons. The three independent variables are each statistically significant with very low p-values. The coefficients are all close to one. For predicted strikeout rate, the coefficient is 0.9, indicating a one-percentage point increase in predicted strikeout rate corresponds to a 0.9 percentage point increase in actual strikeout rate. The coefficient on predicted walk rate is also 0.9, so a one percentage point increase in predicted walk rate correlates to a 0.9

percentage point increase in actual walk rate. The coefficient on home run rate is 0.9, meaning a one percentage point increase in predicted home run rate indicates a 0.9 percentage point increase in actual home run rate. The constants are all statistically significant and positive, suggesting Szymborski underestimates each of these three metrics.

The adjusted R-squared values are all lower than 0.7. The adjusted R-squared is highest on the strikeout rate model at 0.674, which continues the trend of strikeout rate being a bit easier to predict than walk rate and home run rate, and means Szymborski's strikeout rate predictions account for 67.4% of the variance in real strikeout rate. On the walk rate model, the adjusted R-squared is 0.531, so only ZiPS' predicted walk rate accounts for 53.1% of the variance in real walk. The adjusted R-squared on the home run rate model is similar, 0.518, meaning ZiPS' predicted home run rate accounts for 51.8% of the variance in real home run.

I calculated the root mean square error on Szymborski's predictions, then used the MLB on MLB models from section 8.1 to make predictions on the same players from the 2015-2016 seasons, and calculated the root mean square error for these predictions. Because my MLB on MLB models predict statistics for a three-year window, I used the 2012-2014 seasons to predict metrics for the 2015-2017 seasons, and used the latter three-year window as my 2015 and 2016 predictions. The root mean square error on predicted strikeout rate versus actual strikeout rate for Szymborski's batters predictions is 0.174, which is only slightly lower than the root mean square error on predictions with my model, 0.186.

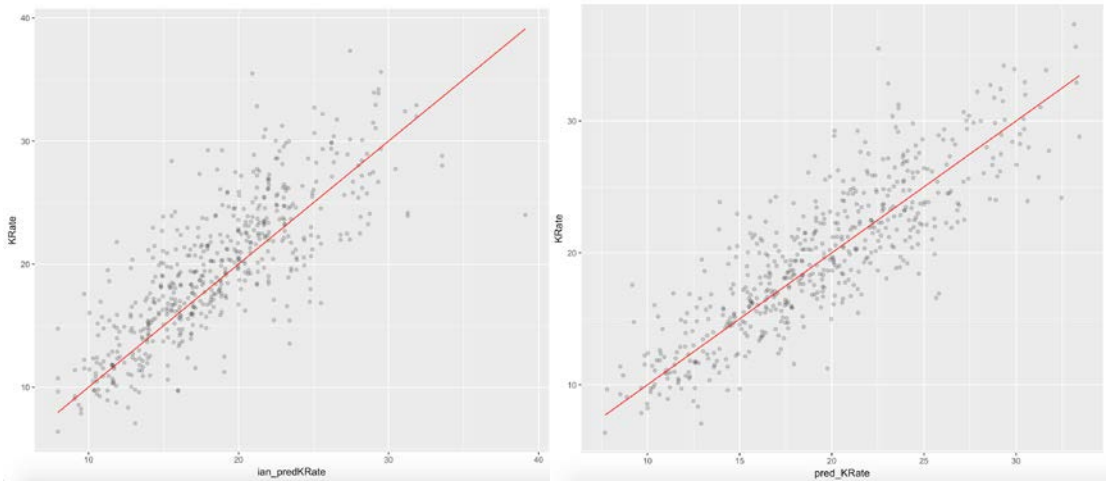

Figure 7

Figure 7 shows my a plot of my predictions and the actual strikeout rates on the left, and ZiPS predictions with the actual strikeout rates on the right. The red line represents what a perfectly linear relationship would look like. Szymborski's predictions appear to center around the red line a bit tighter than mine do, as I see a few more outliers towards the upper-right side of the plot containing my predictions. On predicting walk rate for batters, ZiPS yielded a 0.312 root mean square error, lower than the 0.388 figure my predictions yielded.

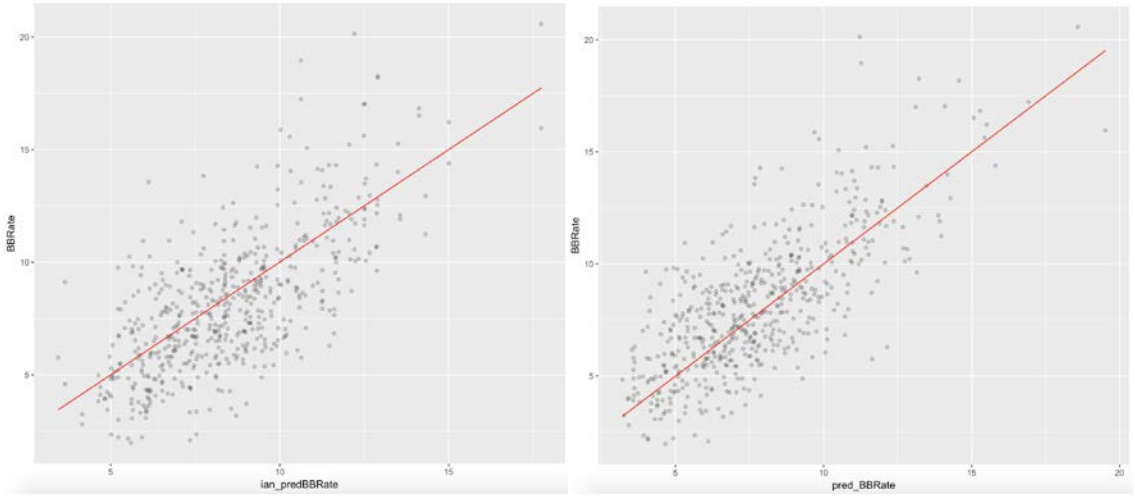

Figure 8

In figure 8, again my predictions are on the left, with the ZiPS predictions on the right. Both Szymborski and I struggled a bit more to fit our predictions around the red line this time. The root mean square error on home run rate predictions for batters was 0.281 for ZiPS and 0.107 for the MLB on MLB model. My home run rate predictions significantly outperformed Szymborski's.

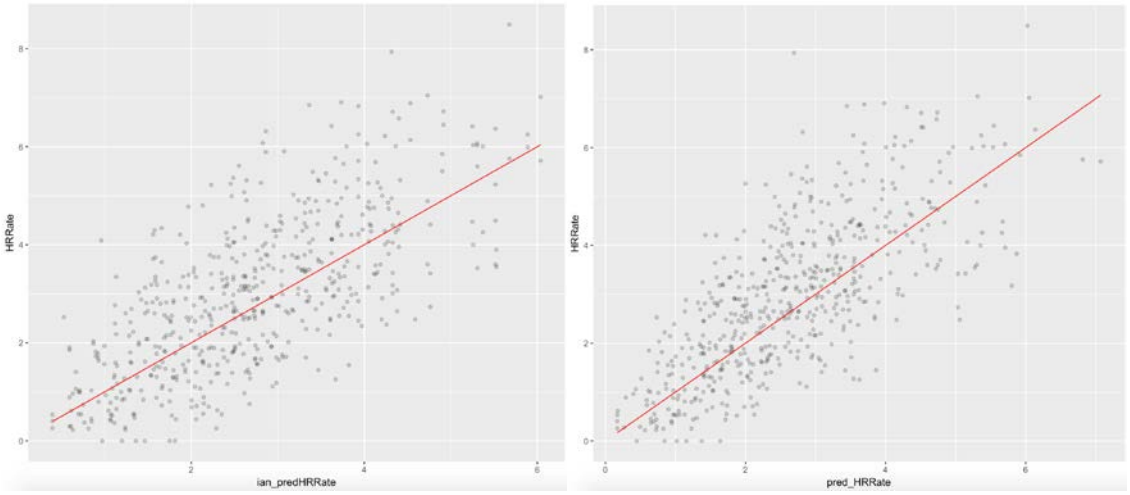

Figure 9

Figure 9 shows my home run rate predictions on the left, and Szymborski's on the right. It appears that both of us have points significantly to the left and above of the red line. ZiPS has a few outliers in the 7-8% range that likely spiked the root mean square error; my max-predicted home run rate was 6.04% (Chris Davis).

	Dependent variable:		
	KRate	BBRate	HRRate
	(1)	(2)	(3)
pred_KRate	0.869***		
	(0.033)		
pred_BBRate		0.686***	
		(0.040)	
pred_HRRate			0.628***
			(0.068)
Constant	3.511***	2.512***	1.166***
	(0.683)	(0.319)	(0.174)
Observations	682	682	682
R2	0.511	0.304	0.113
Adjusted R2	0.511	0.303	0.111
Residual Std. Error (df = 680)	4.068	2.022	1.029
F Statistic (df = 1; 680)	711.658***	297.674***	86.345***
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 9

Table 9 shows the results of the same regressions with pitchers data from Szymborski and Lahman. It appears that pitcher performance is more difficult to predict for Szymborski. The coefficients on the independent variables are all statistically significant. In the strikeout rate model, the coefficient is equal to 0.9, so a one-percentage point increase in predicted strikeout rate correlates with a 0.9 percentage point increase in actual strikeout rate. The coefficient on predicted walk rate is equal to 0.7, so a one-percentage point increase in predicted walk rate only corresponds to a 0.7 percentage point increase in actual walk rate. On predicted home run rate, the coefficient is equal to 0.6, meaning a one percentage point increase in predicted home run rate correlates to a 0.6 percentage point increase in actual home run rate. The constants are also all statistically significant. They are greater than one in the strikeout rate and walk rate

models, indicating Szymborski underestimated these metrics in his predictions, while the home run rate model constant is less than one, suggesting Szymborski overestimated home run rate in his predictions.

The adjusted R-squared value on the strikeout rate model is 0.511, so Szymborski's predicted strikeout rate accounts for 51.1% of the variation in real strikeout. On the walk rate model, the adjusted R-squared is 0.304, meaning ZiPS' walk rate predictions account for 30.4% of the variance in actual walk rate. On the home run rate model, the adjusted R-squared was 0.111, so only ZiPS' home run rate predictions only account for 11.1% of the variance in actual home run rate.

Again, I used my models from section 8.1 to make my own predictions for strikeout rate, walk rate, and home run rate for pitchers in the 2015 and 2016 seasons for a comparison against the ZiPS predictions. The root mean square error for on Szymborski's predicted strikeout rate vs. actual strikeout rate is 0.197, whereas my predictions yield a root mean square error of 0.205, so Szymborski once again narrowly edges me in predicting strikeout rate.

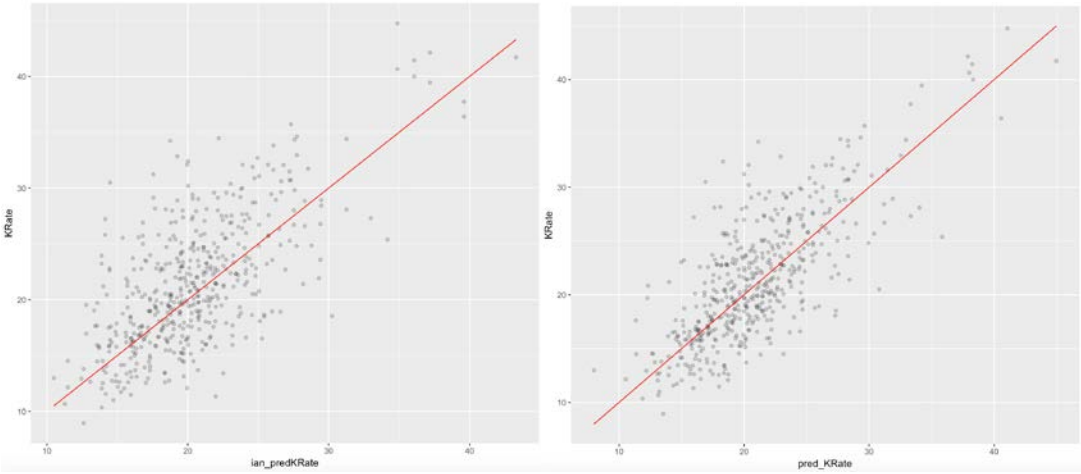

Figure 10

Figure 10 shows my strikeout rate predictions versus the actual strikeout rates on the left, and ZiPS' predictions versus the actual strikeout rates on the right, with the red line representing a perfectly linear relationship. ZiPS predictions appear to fit tighter around the red line than mine. Szymborski's root mean square error for predicted walk rate on actual walk rate is 0.382, and mine is 0.438. Again, ZiPS outperforms my model in predicting walk rate.

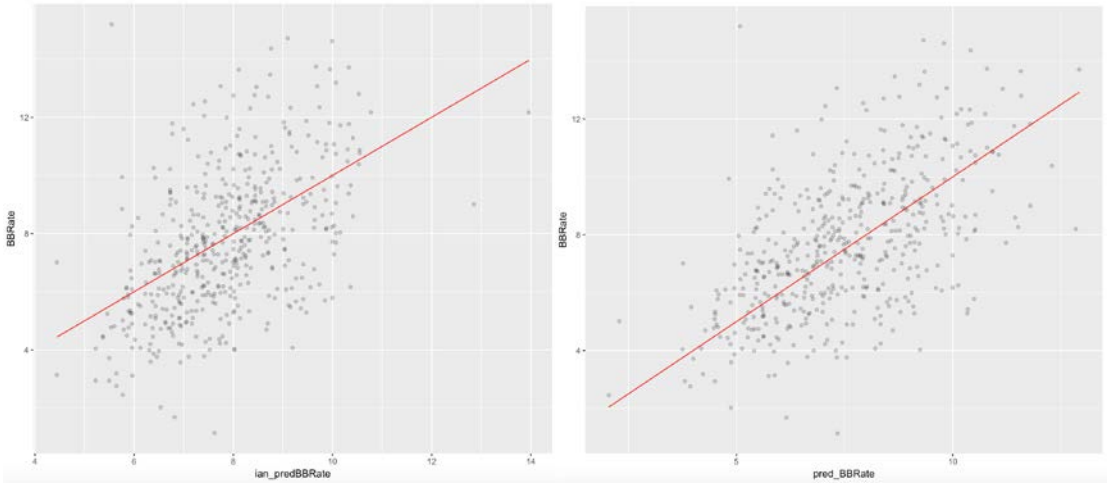

Figure 11

Figure 11 shows my predictions for walk rate on the left, and Szymborski's on the right. We both struggled to predict pitcher walk rate with relatively high root mean square error values, and the plots support this. The root mean square error on ZiPS' predicted home run rate versus actual home run rate is 0.16, and on my model it is 0.309; this time Szymborski predicts home run rate much more effectively than me.

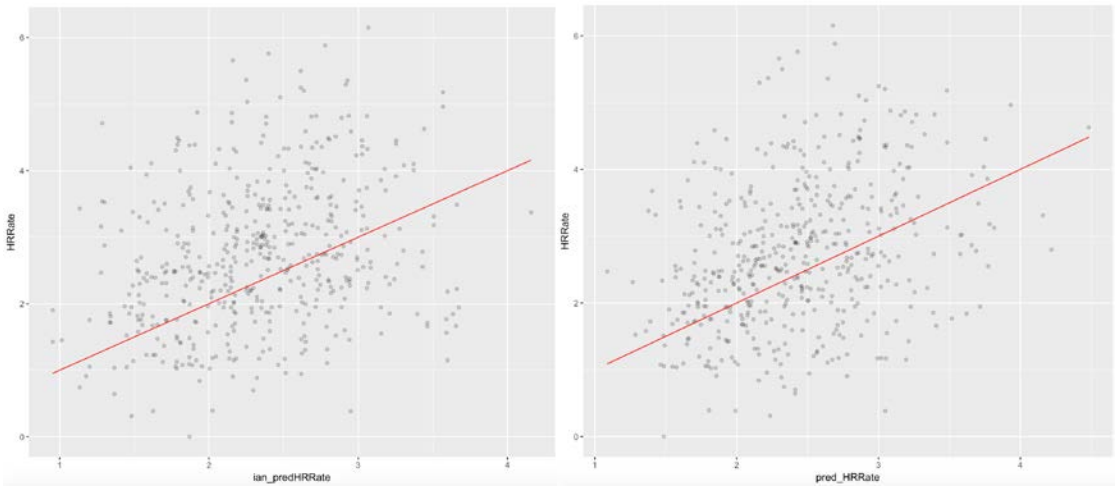

Figure 12

Again, my predictions are on the left in Figure 12, while the ZiPS predictions are on the right. It appears that my model underestimated too many home run rates; my highest prediction was 4.16% (Paul Clemens), and second highest was only 3.67% (Jason Motte).

9. Conclusion

Variation in player performance is apart part of what makes sports interesting. A player like J.D. Martinez can be released by one of the smartest organizations in sports, the Houston Astros, only to end up on a different team for a minimum salary and become one of the best hitters in baseball. Even the best front office executives in the world can be stumped when it comes to evaluating players in an open market like free agency. Theo Epstein is the current General Manager of the Chicago Cubs, and an executive who will one day enter the Baseball Hall of Fame. Even he is not perfect, infamously valuing Carl Crawford so highly after the 2010 season that he signed him to a 7-year/$142 million contract, only to see him traded in a salary-dump trade less than two years into his contract, and only after Epstein was fired by the Boston Red Sox.

My analysis on player predictions based on my MLB on MLB models and Dan Szymborski's ZiPS projection system reveals that predicting performance in general is

extremely difficult. Although the adjusted R-squared values on these models were much greater than on my foreign/amateur league models, they were still not particularly close to one. ZiPS' predictions beat my own predictions based on my models using solely past player performance for each metric except batter home run rate, although my root mean square errors were close to those that ZiPS' predictions yielded for the most part. ZiPS is a commonly used and cited projection system, published annually on Fangraphs.com, one of the premier sources for statistical analysis in baseball. Yet even a projection system as highly thought of as ZiPS is far from perfect.

When it comes to evaluating how a player's performance will translate from the NCAA, KBO, or NPB to the MLB, there appear to be too many factors at hand. For players coming from the NPB and KBO, many of these are native to Japan and South Korea, and may be experiencing western culture for the first time, thrust into not only what could be an uncomfortable environment, but also tasked with facing the toughest baseball competition in the world for the first time. Professional teams in all sports have acknowledged the effect an athlete's mental state can have on player performance with the recent trend of teams hiring sports psychologists. As for players coming from college, the MLB has a steep learning curve, and using the three-year window restriction is more complicated when looking at college players because I limit the seasons before a player enters his prime in his late 20s most of the time since the transition happens at a young age. It can take years for a player to break out and reach his full potential.

What my project reveals is that there is a lot of noise inherent in player performance in baseball. For every foreigner who struggles to perform at a high level in their first few seasons playing in the United States, there are others who thrive in the

same situations. A player might be hitting his physical and athletic peak in his first year playing Major League Baseball, and would have seen a jump in performance had he stayed in playing in the NCAA or overseas. Injuries are also unpredictable, and attempting to play through them can distort player evaluations when using statistics. This makes predicting the performance of players adjusting to completely new environments both on the field and off of it a tricky task.

My project reveals a couple of noteworthy trends that could be worth further study. The first is that age seemed to affect pitcher performance more than batter performance, which fits in lines with the findings of Fair (2008). Another pattern I observed was that strikeout rates are much easier to predict to than walk rates and home run rates. Lastly, batter performance, based on these metrics, appears to be generally easier to predict than pitcher performance.

Perhaps as statistical analysis in baseball continues to evolve, newly developed metrics could paint a clearer picture on how performance will translate. It could be that breaking down a batter's swing plane or a pitcher's arm action, a more mechanical of evaluation a player, is a better way to predict performance in a different competitive environment. Taking a "process is greater than results" approach, it could be that hard-hit rates for both batters and pitchers are a better true indicator of quality. Either way, statistical analysis, in its current form and through the metrics studied in this paper, is not particularly predictive of MLB performance when coming from a different league.

Despite this, there is still evidence that teams are taking performance in other leagues into consideration when evaluating potential signees. Just this past winter, the St. Louis Cardinals signed Miles Mikolas, a pitcher who threw 91.1 innings across the 2012-

14 seasons in the MLB and compiled a mediocre 5.32 ERA before thriving in the NPB in the 2015-17 seasons. Mikolas parlayed a stellar 2.18 ERA and 5.48 strikeout-to-walk ratio in 424.2 innings into a 2-year/$15.5 million contract and another shot in the MLB. Unless the Cardinals uncovered some sort of mechanical change that Milokas made in recent years, it is hard to argue that their front office did not look at his statistics and liked what they saw. So yes, statistical analysis does still matter when evaluating players in other leagues, and major league teams continue to acknowledge that. It is a matter of understanding that even strong player performance like that of Mikolas often contains more noise than signal.

APPENDIX

All coding for this project was done in R. These are descriptions of the various R files:

Extra Data.R

This file contains the wrangling for the supplementary data I used. Since the original sources for KBO and NPB data only went through the 2008 season, and a significant portion of the players to transition from these leagues to the MLB made the transition after 2008, I added data on post-2008 players from baseballreference.com. I read the various baseballreference.com files, assign name and league dummy variables, and combine them into four separate files: “extra nbp batting.csv”, “extra npb pitching.csv”, extra “kbo batting.csv”, and “extra kbo pitching.csv”. These files are exported and used in “Thesis Data.R”.

Thesis Data.R

This file contains all of the data manipulation from the raw MLB, NCAA, NPB, and KBO files as well as the files from Extra Data.R. I read in the files, select the variables needed for my project, rename the variables so that they are consistent across the different data sheets, join master sheets with the sheets containing statistics, assign various league dummy variables, calculate an estimated age variable for the MLB, NPB, and KBO files, combine the sheets into one for batters and one for pitchers, including the supplementary data exported from “Extra Data.R”, create player variables containing first and last names, construct my three-year observation windows, tag the final non-MLB seasons and corresponding ages, filter out some duplicates, collapse the data so that each observation contains a player’s non-MLB and MLB statistics across the three-year windows being studied, and export these final data sheets for use in “Thesis Regressions.R”. This file also contains the wrangling for my MLB on MLB analysis (section 3), which involved creation of a player variable, applying Carleton’s 170 rule, creation of a cumulative years variable and corresponding filtering, splitting the data into pre (seasons 1-3) and post (seasons 4-6) sheets, and the exporting of these sheets for use in “MLB on MLB Analysis.R”

Thesis Regressions.R

This file contains the code for the foreign/amateur regressions in Chapter 6. I read in the exported files from “Thesis Data.R”, calculate the strikeout rate, walk rate, and home run rate metrics, apply Carleton’s 170 rule, write code for descriptive stats, plots, strikeout rate/walk rate/home run rate regressions for both batters and pitchers, tables, and plate appearances/batters faced models discussed in Chapter 7.

MLB on MLB Analysis.R

This file contains the code for my MLB on MLB analysis done in Chapter 8. It begins with reading in the files, applying Carleton’s 170 rule, calculating strikeout rate, walk

rate, and home run rate, then modeling and coding tables. I eventually use these models to make predictions for use of comparing against my models' predictive power with the ZiPS' projections. I load in the "ZiPS Batting Predictions 2015 and 2016.csv" and "ZiPS Pitching Predictions 2015 and 2016.csv" files exported from "ZiPS Analysis.R", load in the Sean Lahman data, rename some of Lahman's variables, join the Lahman statistics with the Lahman master sheet, collapse the data so that each observation contains a player's statistics for the 2012-2014 seasons, calculate the strikeout rates/walk rates/home run rates to be used as inputs for my MLB on MLB model predictions, join this data with the ZiPS data, and make predictions for 2015 and 2016 strikeout rate, walk rate, and home run rate for batters and pitchers, then calculate the root mean square error for both my predictions and the ZiPS projections, and compare graphs.

ZiPS Analysis.R

This file contains the code for analysis of Dan Szymborski's ZiPS projection system. I read in the ZiPS files for the 2015 and 2016 seasons, combine the seasons into one sheet for batters and one sheet for pitchers, calculate his predictions for strikeout rate, walk rate, and home run rate, re-split the data into separate seasons, load in the Lahman data, rename Lahman's variables, join the Lahman sheets with Lahman's master sheet, calculate strikeout rate/walk rate/home run rate with the Lahman data, split the Lahman data into 2015 and 2016 sheets, combine these with ZiPS 2015 and 2016 sheets, then re-combine the 2015 and 2016 sheets for batters and pitchers, model with these sheets, calculate root mean square error to evaluate the accuracy of ZiPS, and export these final sheets containing ZiPS predictions for use in "MLB on MLB Analysis.R".

BIBLIOGRAPHY

"Baseball Reference." *Sports Reference LLC*, 2000-2017. Accessed September 26, 2017. https://www.baseball-reference.com/.

"The Baseball Cube." 2003-2017. Accessed October 6, 2017. http://www.thebaseballcube.com/about/

"Baseball Statistics and Analysis." 2006-2017. Accessed September 21, 2017. http://fangraphs.com

Albright, Jim. "Guru's Clubhouse." 1998-2017. Accessed October 3, 2017. http://baseballguru.com/jalbright/

Carleton, Russell. "525,600 Minutes: How Do You Measure a Player in a Year?." Fangraphs, April 20, 2011. Accessed November 14, 2017. https://www.fangraphs.com/blogs/525600-minutes-how-do-you-measure-a-player-in-a-year/

Davenport, Clay. "A Repository for Stats I Still Care About." 2011-2017. Accessed September 21, 2017. http://claydavenport.com/

Druschel, Henry. "A guide to the projection systems." Beyond the Box Score, February 22, 2016. Accessed March 9th, 2018. https://www.beyondtheboxscore.com/2016/2/22/11079186/projections-marcel-pecota-zips-steamer-explained-guide-math-is-fun

Fair, C. Ray. "Estimated Age Effects in Baseball." Journal of Quantitative Analysis in Sports, Vol. 4, Issue 1, 2008. Accessed October 21, 2017.

Franks, M. Alexander, Alexander D'Amour, Daniel Cervone, Luke Bornn, "Meta-analytics: tools for understanding the statistical properties of sports metrics." Journal of Quantitative Analysis in Sports, Vol. 12, Issue 4, 2016. Accessed September 22, 2017.

Keri, Jonah. "The Evolution of the Cuban Baseball Pipeline." Grantland, June 18, 2014. Accessed October 4, 2017. http://grantland.com/the-triangle/the-evolution-of-the-cuban-baseball-pipeline/

Lindbergh, Ben. "What the Red Sox's Rusney Castillo Signing Means for the Player, the Team, and Future Cuban Free Agents." Grantland, August 26, 2014. Accessed October 3, 2017. http://grantland.com/the-triangle/boston-red-sox-rusney-castillo-future-of-cuban-free-agency/

McCracken, Voros. "V÷R÷S' BASEBALL ANALYSIS PAGE: Defense Independent Pitching Stats, Version 2.0 Formula." Futility Infielder, January 17, 2002. Accessed March 2, 2018. http://www.futilityinfielder.com/dipsexpl.html

Petti, Bill. "Pitcher Aging Curves: Introduction." Fangraphs, April 30, 2012. Accessed October 21, 2017. https://www.fangraphs.com/blogs/pitcher-aging-curves-introduction/

Swartz, Matt. "The Recent History of Free Agent-Pricing." Fangraphs. July 11, 2017. Accessed October 13, 2017. https://www.fangraphs.com/blogs/the-recent-history-of-free-agent-pricing/

Weinberg, Neil. "Basic Principles of Free Agent Contract Valuation." Fangraphs, January 14, 2016. Accessed October 13, 2017. https://www.fangraphs.com/library/basic-principles-of-free-agent-contract-evaluation/

Zimmerman, Jeff. "Component Changes in New Hitter Aging Curves." Fangraphs, January 15, 2014. Accessed October 20, 2017. https://www.fangraphs.com/blogs/component-changes-in-new-hitter-aging-curves/