May 2014

Improving the FIP Model

Joseph Patrick Flanagan
Worcester Polytechnic Institute

Project Number: MQP-SDO-204

Improving the FIP Model

A Major Qualifying Project Report

Submitted to The Faculty

of

Worcester Polytechnic Institute

In partial fulfillment of the requirements for the

Degree of Bachelor of Science

by

Joseph Flanagan

April 2014

Approved:

Professor Sarah Olson

Abstract

The goal of this project is to improve the Fielding Independent Pitching (FIP) model for evaluating Major League Baseball starting pitchers. FIP attempts to separate a pitcher's controllable performance from random variation and the performance of his defense. Data from the 2002-2013 seasons will be analyzed and the results will be incorporated into a new metric. The new proposed model will be called jFIP. jFIP adds popups and hit by pitch to the fielding independent stats and also includes adjustments for a pitcher's defense and his efficiency in completing innings. Initial results suggest that the new metric is better than FIP at predicting pitcher ERA.

Executive Summary

Fielding Independent Pitching (FIP) is a metric created to measure pitcher performance. FIP can trace its roots back to research done by Voros McCracken in pursuit of winning his fantasy baseball league. McCracken discovered that there was little difference in the abilities of pitchers to prevent balls in play from becoming hits. Since individual pitchers can have greatly varying levels of effectiveness, this led him to wonder what pitchers did have control over. He found three that stood apart from the rest: strikeouts, walks, and home runs. Because these events involve only the batter and the pitcher, they are referred to as "fielding independent."

FIP takes only strikeouts, walks, home runs, and innings pitched as inputs and it is scaled to earned run average (ERA) to allow for easier and more useful comparisons, as ERA has traditionally been one of the most important statistics for evaluating pitchers. To operate on this scale, FIP includes a yearly calculated constant such that league-average FIP equals league-average ERA. In this study I have sought to improve upon the simple FIP model with a new metric, jFIP. jFIP is very similar to FIP, operating on the same fundamental principles, but with slightly modified data input. The new model takes pop ups and hit by pitch as inputs, in addition to the existing ones of strikeouts, walks, home runs, and innings pitched. It is also adjusted based on how well a pitcher managed to field his position. In order to create this model, data from 2002-2013 were examined and fed into the FIP equation, with the strikeouts input being replaced by strikeouts plus pop ups and the walks input replaced by walks plus hit by pitch. Then the league average jFIP was calculated for each year and used to compute the constant, which is close to the one used by FIP.

A matched pairs t test was used to evaluate the metric's accuracy at modeling ERA. A matched pairs t test is used when you have two samples of data for the same subjects, often a before and after type scenario. In this case, the "before" is simply FIP before I did anything to modify it, while the "after" is jFIP. The test provides a value, t, that can then be checked against a table of critical values for the proper signficance level and degrees of freedom. If t exceeds the critical value, then the hypothesis that jFIP is no better than FIP can be rejected in favor the hypothesis that it is indeed better. For samples of this size (n>60), the critical threshold for the t statistic is approximately 2. After adding the new data inputs, none of the years in this study had t values exceeding the critical value, but after also adjusting for pitcher defense, jFIP 2 was found to be statistically significantly better than FIP at predicting ERA in 5 out of the 11 years, with one more year just missing the established threshold.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Baseball

Fielding Independent Pitching, or FIP, is a pitching metric developed by Tom Tango [4] and based off of research by Voros McCracken. It seeks to isolate a pitcher's performance from the contributions of the defense behind him and from the effects of random chance. McCracken's breakthrough discovery that pitchers had little ability to prevent balls in the field of play from becoming hits went against what had been previously assumed by baseball fans and analysts. If pitchers could not control whether balls in play dropped for hits, then what exactly separated the good pitchers from the bad? According to FIP, it is the pitcher's ability to record strikeouts, avoid walks, and limit home runs. Despite allowing hits on balls in play at a comparable rate to lesser pitchers, a good pitcher can strike out enough batters that there are simply fewer balls put in play against him, and therefore fewer total hits against him. Similarly, a good pitcher can use his command of the strike zone to limit the number of walks he allows. A walk is generally worse for the pitcher than a ball in play, as there is no chance to put out a batter at first once he has walked, so the ability to prevent free baserunners is important. And a home run is the best case scenario for any batter in any situation, so it is clear that being able to keep the ball in the park will help a pitcher in the long run.

1.2 Important Terms and Acronyms

- IP = innings pitched. This is how many innings a pitcher threw in a given year. Outs count as partial innings, such that a pitcher who pitches the first five innings of a game and proceeds to only record one out in the sixth before he is replaced is credited with 5.1 innings pitched, though this is read aloud as "five and one third."
- K% = strike out rate. This is the percentage of opposing batters that a pitcher strikes out. Higher is better.
- BB% = walk rate. This is the percentage of opposing batters a pitcher walks. Lower is better.
- HR% = home run rate. This is the percentage of opposing batters a pitcher allows a home run to. Lower is better.

In addition to the stats above, the following will also prove important throughout the course of this project:

- PU = pop up. A pop up is a high, shallow fly ball, and pop ups are generally automatic outs. Batted ball groupings do not really have precise definitions, so this number may change based on the data source. All batted ball data in this report comes from www.fangraphs.com
- HBP = hit by pitch. A hit batsman gets first base for free. They are essentially walks, except that runners may not advance after a HBP.
- LOB% = left on base. This is the percentage of opposing baserunners that a pitcher strands on base.
- WP = wild pitch. This is the number of wild pitches the pitcher threw. A wild pitch is a pitch that the catcher can not handle cleanly due to a bad throw from the pitcher. Note that a wild pitch may only be recorded if one or more runners advance on the play.
- PA = plate appearance. Every time a batter completes his time at bat (barring the inning ending prematurely on a baserunning out), it is recorded as a plate appearance. All of a player's counting stats (at bats, sacrifices, walks, etc.) are included in his plate appearance total.
- AB = at bat. An at bat is any plate appearance that doesn't end by walk, hit by pitch, sacrifice fly, or sacrifice bunt.
- SF = sacrifice fly. A sacrifice fly is a fly ball caught with fewer than two outs that allows a runner on third to tag up and score. Sacrifice flies are counted as a plate appearance, but not as an at-bat. It is described here only for use in the BABIP formula.
- BABIP = batting average on balls in play. This is the percentage of balls in play allowed by the pitcher that fall in for hits. Home runs are not considered "in play" because the defense does not have a chance at making an out.

$$BABIP = \frac{H - HR}{AB - HR - K + SF} \tag{1.1}$$

- ER = earned runs. An unearned run is any run that, in the official scorer's judgement, would not have scored without an error or passed ball. Earned runs are all those runs that are not unearned (i.e. score mostly because of the pitcher) and make up the vast majority of runs scored.
- ERA = earned run average. This is the number of runs, on average, that a pitcher allows per nine innings pitched, not counting runs that scored due to a fielder's error.
- FIP = fielding independent pitching. This is what a pitcher's ERA "should" have been, based on his K, BB, and HR rates and assuming league average results for BABIP, LOB%, and sequencing.

1.3 Deriving FIP

The FIP formula is as follows:

$$\frac{13 * HR + 3 * BB - 2 * K}{IP} + C \tag{1.2}$$

C is a constant, calculated each year after the season, that is used to bring FIP in line with the ERA scale, such that league-average FIP = league-average ERA.

The FIP equation is derived from linear weights, introduced by Pete Palmer in his 1984 book, "The Hidden Game of Baseball" [3]. Linear weights seek to assign a value to every offensive event in baseball. It is a run estimator that attempts to predict how many runs a team "should" score. Not every baseball season sees the same number of runs scored, though. Due to rule changes, changing rule enforcement, equipment, illegal supplements, or any number of things, the offensive "environment" of MLB is not always the same. In low run environments, such as 1968, which saw pitchers dominate after a redefinition of the strike zone, runs were difficult to come by, and therefore each run was more valuable, as it took fewer of them to win a game. If it only takes 3 runs to win a game on average, each of those runs is more valuable than a run scored during a period where it usually takes 6 runs to win. It is important to note that the run values in linear weights are given compared to the league average. A team with 0 runs above average (RAA) would be considered a league-average offense.

Palmer's original linear weights formula is as follows:

$$RAA = (.47 * 1B + .78 * 2B + 1.09 * 3B + 1.4 * HR + .33 * (BB + HBP) - .27 * (AB - H)) \quad (1.3)$$

The most valuable offensive event is a home run, followed by a triple, double, and single in that order. A walk or a hit by pitch (HBP) are worth slightly less than a single, as each advances the batter to first like a single, but in either case runners generally do not score. A single will score any runner on third and has somewhere around a 50% chance to score a runner on second. Walks and HBP can only score a runner if the bases are already loaded. The term (AB-H) is a simple representation for outs, and thus is assigned a negative value. Not making an out always has positive value for the batting team, while making an out always carries negative value.

To derive the FIP equation [5], Tom Tango classified the components according to the pitcher's control over them. Strikeouts, walks, and home runs were considered under a pitcher's control, while singles, doubles, triples, and outs were not (HBP were apparently ignored, likely considered marginal due to their infreqeuncy). The former 3 were to become components of FIP, while the rest were lumped together as balls in play (BIP). The average value for any BIP was calculated as -.03 runs.

From equation 1.3, the value of an out is given as -.27 runs. It makes sense, then, the value of a strikeout, specifically, is also -.27 runs. So simplifying the equation into the FIP components and BIP gives

$$RAA = (-.27 * K - .03 * BIP + .33 * BB + 1.4 * HR) \quad (1.4)$$

To convert runs above average to simply runs, .12 runs per plate appearance are added to each event. There are an average of 38.5 plate appearances and 4.62 runs per team per game, which works out to a baseline of .12 runs per plate appearance.

$$Runs = (-.15 * K + .09 * BIP + .45 * BB + 1.52 * HR) \quad (1.5)$$

Since FIP does not consider balls in play, we will subtract .09 runs per plate appearance from every event so that the coefficient of BIP is 0. The .09 runs per plate appearance becomes its own term.

$$Runs = (-.24 * K + .36 * BB + 1.43 * HR) + .09 * PA \quad (1.6)$$

An average of 38.5 plate appearances per game gives

$$Runs = (-.24 * K + .36 * BB + 1.43 * HR) + .09 * 38.5 \quad (1.7)$$

9 innings per game leaves

$$Runs = (-.24 * 9 * K/IP + .36 * 9 * BB/IP + 1.43 * 9 * HR/IP + .09 * 38.5) \tag{1.8}$$

$$Runs = \frac{-2.16 * K + 3.24 * BB + 12.87 * HR}{IP} + 3.47 \tag{1.9}$$

Lastly, to convert from total runs to earned runs, multiply by .923. earned runs =

$$\frac{-1.99 * K + 2.99 * BB + 11.87 * HR}{IP} + 3.20 \tag{1.10}$$

So we can see the constant in the FIP equation exists as an equalizer for pitchers' balls in play. Whether the offensive environment is particularly high or low, the effects of strikeouts, walks, and home runs on run scoring do not change too drastically. What tends to change more is the value of balls in play. In a high offensive environment where it's easier to score, the value of a ball in play increases because it is much more likely to go for a hit than a ball in play during a low run-scoring period. Walks, strikeouts, and home runs still have much of the same interplay between them, so the constant provides context for how the pitcher is assumed to have fared on his balls in play. It is also true that the constant does not change significantly from year-to-year, and so adjusting the constant slightly is more preferable than recalculating the coefficients for the current run environment. FIP was created with simplicity partly in mind, and the usage of the constant accomplishes this while making sense within a baseball context.

1.4 The Problems with FIP

Like ERA, FIP seeks to determine how effective a pitcher has been, but in a very different way. Sequencing is the order in which various baseball events happen. For example, say a pitcher throws one inning and allows a home run and two walks while striking out three. One way that this could be sequenced is BB, K, K, BB, HR, K. Because this sequence has two men on base when the home run occurs, the pitcher gives up 3 earned runs in this inning. Now let's try it another way and say the inning instead goes HR, K, K, BB, BB, K. Now the first batter hits a home run for only one run, and the other two men who reach base are stranded there at the end of the inning. This sequence produces a much improved result of only 1 earned run. For sequencing to be under a pitcher's control would imply that pitchers have the ability to "bear down," meaning they can strike out batters with greater frequency when in a jam or following a home run. This has not been demonstrated to any convincing degree and so sequencing is considered to fall under the umbrella of random chance.

Consider also the case of two batted balls to the third baseman. The first is a weakly hit ground ball rolling down the foul line that the batter has time to beat out. The second is a screaming line drive over the third baseman's head that he catches with a perfectly timed jump. In which scenario has the pitcher been more effective? Most fans would agree that he was more effective on the ground ball, where he managed to keep the hitter off-balance and induce very weak contact, whereas the line drive was squared up and would have been an easy double if it were just a few inches higher or the third baseman a few inches shorter. In the first case, the pitcher has been unlucky that such weak contact produced a hit anyway. In the second example, the pitcher has been fortunate, because high quality contact still managed to produce an out.

FIP attempts to strip out these elements of sequencing and luck, as it holds that these are not things that an individual pitcher has control over. Instead, it completely ignores sequencing, preferring to simply count up these events, and assumes average luck when it comes to batted balls dropping for

hits. Because it does not include these elements of randomness, a pitcher's FIP tends to remain more stable year-to-year than his ERA, which can fluctuate wildly with just a few cases of bad luck strung together.

This is where I think FIP does not go quite far enough, though. I believe there are other factors under a pitcher's control that can and should be accounted for in FIP, such as certain batted ball ratios and the ability to field the position and hold the running game. In addition, I believe using innings pitched as the denominator tends to underrate the good pitchers and overrate the bad ones, since the best pitchers in baseball will face fewer batters per inning than the average, and the worst will face more batters.

Chapter 2

The Data

The sample of pitchers used in this study is all qualified starters for each year in the span 2002-2013. A starting pitcher qualifies for the ERA title if he pitches at least one inning for every game his team plays, or 162 innings with no rainouts or other cancellations. Generally, there are between 85 and 95 starting pitchers each year, spread over 30 teams, who qualify. I chose the period 2002-2013 because that is the time for which there is accurate batted ball data and detailed fielding logs, which will be needed to compute jFIP.

To obtain this sample, I created a custom leaderboard at www.fangraphs.com [1] in order to pick and choose which stats I wanted to examine. I selected the first year of the desired range, 2002, and added columns for any relevant statistics I wanted (of which the important ones are detailed below). Then I imported the data into Microsoft Excel. I repeated this process 11 more times until I had 12 spreadsheets total, one for each year of the study. These were then called into Matlab using the "xlsread" command.

The mean of a set of numbers is simply the average. To calculate the mean, all values in the set are summed and then divided by the total number of values. The mean ($\bar{X}$) of 4, 8, and 9 is (4+8+9)/3 = 7.

Standard deviation (s_x) is a measure of how close the data points tend to be to the mean. The lower the standard deviation, the more clustered the values are and the closer those values generally are to the mean. Standard deviation is calculated by taking the square root of the sum of the squared differences between the actual values and the mean.

$$s_x = \sqrt{\sum_{i=1}^{n} (\bar{X} - x_i)^2} \tag{2.1}$$

Let's examine Table 2.1, which contains statistics on all the qualified starters from 2002. The mean strikeout (K) rate was 17%, with the best being about twice as good as that, and the worst being about twice as bad. Walks (BB), too, see a similar relationship, with a mean of 7.6% and the best and worst about twice as low and twice as high. Home runs (HR) have a mean rate of 3.4% and a small standard deviation of just under 1%.

Table 2.2 contains statistics from the 2013 season. The mean strikeout (K) rate here is 20.2%. Even the worst qualified strikeout rate was up to 12.3%, remaining slightly higher than half the mean rate. The maximum strikeout (K) rate is right around 33% again, which is essentially where performance in this area has plateaued. The mean walk (BB) rate was just slightly under 7%, with no one walking more than 12% of batters. Home runs(HR) comprised 3.6% of plate appearances with a standard deviation just under 1% again.

Strikeouts have been on the rise so far this century. Though the top pitchers have remained at a 1 in 3 rate, even the league's worst are striking out 1 in every 8 batters now as opposed to 1 in 12 and the mean has jumped to 1 in 5. Walks and home runs have seen no real change from 2002 to 2013. Pop ups (PU) are slightly down across the board, while wild pitches are slightly up. The mean BABIP is up 8 points, possibly just a side effect of the lower pop up (PU) rate. ERA leaders and trailers can fluctuate wildly each year, but the mean ERA is down .30 points in just over 10 years, and FIP is right alongside it with a .40 drop. None of the standard deviations of these metrics changed meaningfully. Since neither walk nor home run rates changed drastically from 2002-2013, the decrease in FIP must be due to the meaningful increase in strikeouts.

If FIP is scaled to ERA, though, then how can the recent .30 drop in league ERA produce a corresponding .40 drop in FIP, rather than .30? This is due to the split between starting pitchers and relief pitchers. Starting pitchers as a whole have underperformed their FIP (meaning their actual ERAs were higher) for every year of this study. It is only by a small amount, .06 to .07 points worth of ERA per year, but it is consistently present. Necessarily then, relief pitchers have outperformed their collective FIP for every year, by about .12 points of ERA (this discrepancy is about twice that of the starting pitchers' because relievers only account for half as many innings).

Relief pitchers have an inherent performance advantage over starters. They generally pitch one inning at a time at most, and can therefore put more effort into each pitch. It's much easier to throw 95 MPH consistently when you know you will only have to do so for a handful of batters; a starter must pace himself for the long run to ensure he does not tire too quickly. Relief pitchers also do not have to face the same batter more than once in a game, barring extreme circumstances in extra-inning affairs. Starting pitchers will nearly always see the entire lineup twice before being removed, and usually must face at least some of the opposing hitters for a third time. These multiple matchups in a day collectively work in the batter's favor. As the game goes on, the starter tires and loses effectiveness, while the hitters also make adjustments to what they have seen in previous plate appearances. As a rule, the more they see of him, the more major league hitters will be able to "figure out" a pitcher and perform better against him. Therefore, the effectiveness of a pitcher is often maximized in short stints, which is what we can see with relief performance.

All of this serves to explain why relievers generally sustain FIPs lower than starters (maximum effort with each pitch theoretically leads to more K and fewer BB and HR), but it does not answer the question of why relievers see better resuts than expected and why starters see worse. The reason for this is mid-inning pitching changes. When a starter is removed from the game, any current baserunners are considered to be his responsibility. If they score, the starter is the one charged with giving up earned runs, not the reliever who gave up the double that drove them in. Only if that runner who just doubled scores will the relief pitcher have given up a run. If the reliever enters with no outs, then he has no ERA advantage, as the opposing team must put a man on base and score him before recording three outs, just like any inning. But if he enters the game with one or two outs already, now he must only reitre one or two more men before allowing someone all the way around the bases. This is clearly an easier task, akin to changing the number of outs per inning to one or two, and is the reason why relievers will continue to post ERAs lower than their FIPs.

The FIP formula is

$$\frac{(13 * HR + 3 * BB - 2 * K)}{IP} + C, \tag{2.2}$$

where C is a constant. From 2002-2013, the range of 13* HR for a season is from 78 to 598. From 2002-2013, the range of 3*BB in a season is from 27 to 357. From 2002-2013, the range of 2*K in a season is from 92 to 668.

The values of this constant over the time period are summarized in Table 2.3. The mean ($\bar{X}$) of

Table 2.1: 2002 summary statistics

Metric	Min	Max	Mean ($\bar{X}$)	Std. Dev. (s_x)
K%	.088	.323	.170	.044
BB%	.032	.12	.076	.019
HR%	.014	.62	.036	.0096
PU%	.014	.081	.043	.015
HBP%	0	.019	.008	.004
LOB%	.63	.83	.73	.044
WP	0	14	5.28	3.29
BABIP	.232	.326	.282	.021
ERA	2.26	6.15	3.9	.8
FIP	2.24	5.53	4.05	.64

Table 2.2: 2013 summary statistics

Metric	Min	Max	Mean ($\bar{X}$)	Std. Dev. (s_x)
K%	.123	.329	.202	.04
BB%	.037	.118	.069	.018
HR%	.015	.62	.034	.0097
PU%	.009	.067	.033	.012
HBP%	0	.021	.007	.004
LOB%	.633	.839	.743	.041
WP	0	16	6.26	3.51
BABIP	.240	.333	.290	.022
ERA	1.83	5.71	3.61	.72
FIP	2.00	4.79	3.64	.59

Table 2.3: FIP constant (C)

Year	Constant
2002	2.962
2003	3.031
2004	3.049
2005	3.020
2006	3.147
2007	3.240
2008	3.132
2009	3.097
2010	3.079
2011	3.025
2012	3.095
2013	3.048

this constant over the 12 year period is 3.07708 and the standard deviation (s_x) is .0727.

Because of the innings pitched denominator, we will examine the highest and lowest K/9, BB/9, and HR/9 in the 2002-2013 period to determine the total possible range of FIP. Since 2002, the highest HR/9 in a season is 2.08, by Bronson Arroyo (Cincinatti Reds) in 2011. The lowest is .31 HR/9, by Charlie Morton (Pittsburgh Pirates) in 2011. The highest K/9 is 11.89, by Yu Darvish (Texas Rangers) in 2013, and the lowest is 2.13, by Nate Cornejo (Detroit Tigers) in 2003. The highest BB/9 is 5.22, by Ricky Romero (Toronto Blue Jays) in 2012. The stingiest is .43, by Carlos Silva (Minnesota Twins) in 2005.

Plugging these in to the FIP equation (before the constant) yields a minimum value of -2.05 and a maximum value of 4.27. Over the same time period, the lowest seasonal ERA is 1.83 and the highest 6.99. If FIP is to be fit to the ERA scale, then this gives a range of around 2.72 to 3.83 for the FIP constant, which is consistent with the 3.2 or so that is currently in use (Table 2.3).

Chapter 3

Methodologies

In order to create an improved metric based on FIP, I will first need to identify which statistics have a high enough year-to-year correlation with themselves to be considered meaningful. A statistic that is not highly correlated with itself may only serve to introduce the randomness that FIP has been attempting to strip out. Then I will develop a model using linear weights to create a modified version of FIP. The goal is for the new metric, jFIP, to be closer to actual ERA than FIP, or at least to be able to do so for some identifiable subset of pitchers (perhaps it will be more accurate only for those who allow a lot of steals, fly balls, etc.). Sometimes pitchers have large gaps between their ERA and FIP that cannot entirely be explained by bad luck. My new metric seeks to hopefully explain some of those gaps and therefore give a "truer" estimate.

3.1 Correlation

Two basic yet important statistical analyses in this project are correlation and regression.

The correlation coefficient, commonly denoted r, is a measure of the strength of the linear relationship between two variables[2].

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \ , \tag{3.1}$$

where n is the number of data points, $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y, and s_X and s_Y are the sample standard deviations. Mean and standard deviation are defined previously in "The Data."

The value of r can range from -1 to 1. A value of 1 means that there is a perfectly positive linear relationship between the variables. For every increase in X there will be a corresponding increase in Y. A value of -1 means there is a perfectly negative linear relationship between the two. For every increase in X, there is a corresponding decrease in Y. A value of 0 means that the two variables are uncorrelated, with no semblance of a linear relationship.

It can be seen in Figure 3.1 that strikeouts and home runs are the two statistics here most highly correlated with ERA, albeit in opposite directions. This is because home runs are the fastest way to give up runs and strikeouts are the most surefire way to get outs, and therefore, not give up runs. You will notice that the statistics which are good for the pitcher (strikeouts and popups) are negatively correlated with ERA, as more of these will lead to a lower ERA. The other three statistics are all preferable outcomes for the batter, and are positively correlated, leading to a higher ERA.

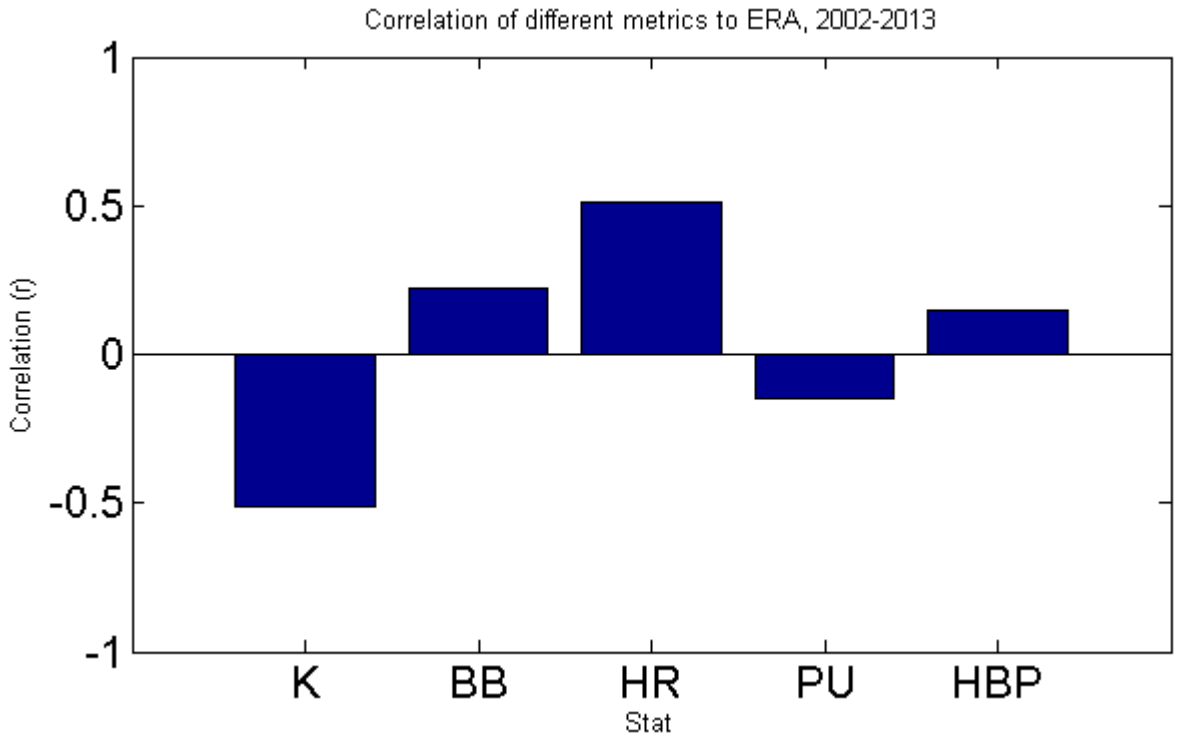

Figure 3.1: Strength of correlation between jFIP stats and ERA

3.1.1 Regression

Regression is a statistical process for estimating the relationship of one variable to another, or to many others[2]. Simple linear regression models the relationship between a dependent variable Y and one explanatory variable, X. With more than one explanatory variable, it is instead called multiple linear regression.

One of the most basic and common types of regression is linear least squares. Regression seeks to find the "best fit" line through the data points, such that the line is as close to as many points as possible. The best fit line should seek to minimize the residuals, or the distance from the line to all of the data points. To do this, the least squares method calculates the residual at every point, squares them and sums them to determine the goodness of fit (the lower the sum, the better the line fits the data). The coefficients of the best fit line's equation are calculated such that this sum is minimized.

For an example of the least squares method in action, let's imagine we have 3 pairs of observations: (4, 6), (8, 13), (19, 25). What is the coefficient of the independent variable in the linear regression equation?

We need to find an equation $y = mx + b$ that minimizes the squares of the residuals

$$R = \sum_{i=1}^{n} (r_i)^2 = \sum_{i=1}^{n} [y_i - (a + b * x_i)]^2$$

To find the minimum, we partially differentiate with respect to a and b and set those equations to 0. This is shown below in Eq. (3.2) and (3.3).

$$\frac{\partial R}{\partial a} = -2\sum_{i=1}^{n}[y_i - (a + bx_i)] = 0 \tag{3.2}$$

$$\frac{\partial R}{\partial b} = -2x_i\sum_{i=1}^{n}[y_i - (a + bx_i)] = 0, \tag{3.3}$$

where n is the number of observations.

These yield the system of normal equations, (3.4) and (3.5):

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i \tag{3.4}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \tag{3.5}$$

Solving the system through the method of elimination gives the least squares estimate for b in (3.6)

$$b = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} \tag{3.6}$$

Then solve the first of the normal equations for a in (3.7)

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} \tag{3.7}$$

This gives b=1.23 and a=1.96, as seen in Figure 3.2.

There are certain baseball metrics that seem like there should be a relationship between the two. For example, LOB% and ERA. Any runner left on base, by definition, did not score and thus cannot possibly count toward ERA. The more runners a pitcher strands, the fewer he should allow to cross the plate. Thus, we would expect to see a negative relationship between the two, as when LOB% increases, ERA should decrease correspondingly. This relationship is illustrated in Figures 3.3 and 3.4.

It is important not to equate correlation with causation. Correlation implies a relationship, but does not necessarily imply that one of the correlated things is causing the change in the other. For example, as we have seen, ERA is negatively correlated with LOB%. But it would be ridiculous to claim that a pitcher's decreased ERA was the cause of him stranding more runners. Stranding runners is a component of good pitching that helps to prevent runs, so if anything the causation would be backwards. It would still be a mistake, though, to assume that the reverse is true, and that higher LOB% is the cause of lower ERA. It certainly might be the case, but there could also be a hidden third variable, whose changing causes both an increase in LOB% and a decrease in ERA.

Another metric we would expect to see have a relationship with ERA is HR%. The more prone a pitcher is to surrendering home runs, the more runs we would expect him to allow, as a home run is at least one run automatically and is always the worst case outcome. Home runs allow all runners to score and can therefore compound other mistakes that put men on base in the first place. We would predict these two to have a positive relationship with each other, as allowing more home runs should lead to a higher ERA. Indeed, Figures 3.5 and 3.6 show this expected trend.

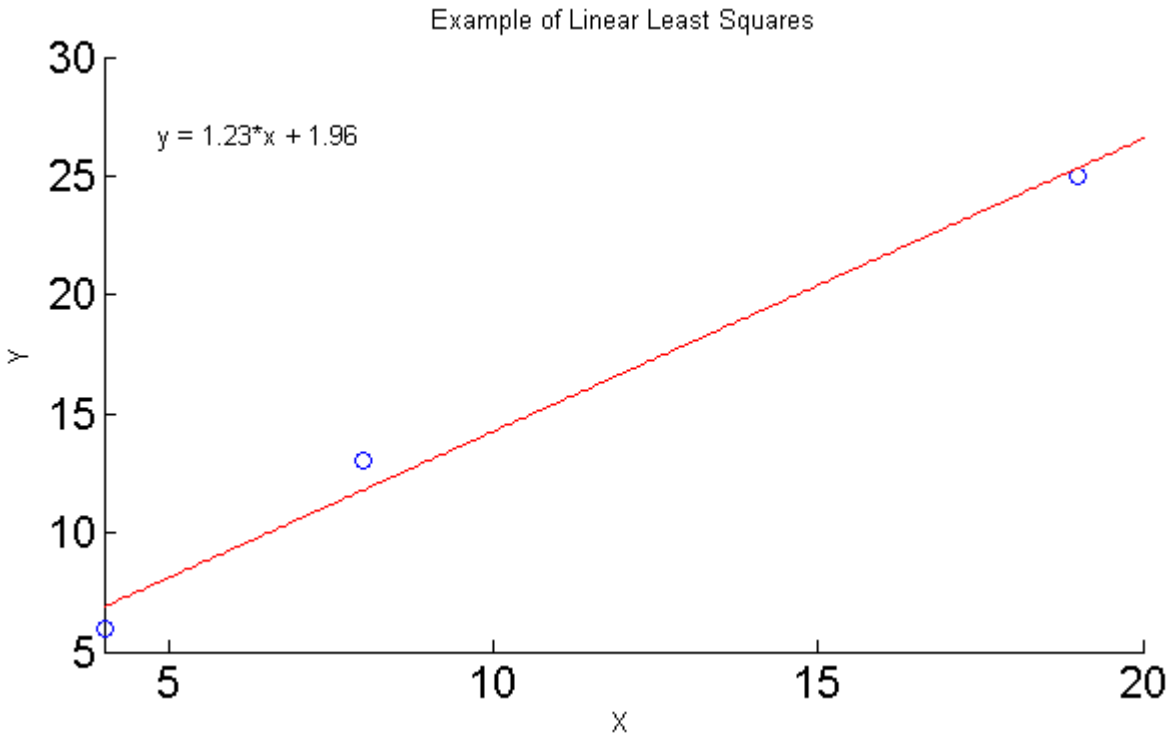

Figure 3.2: A MATLAB example of linear least squares regression

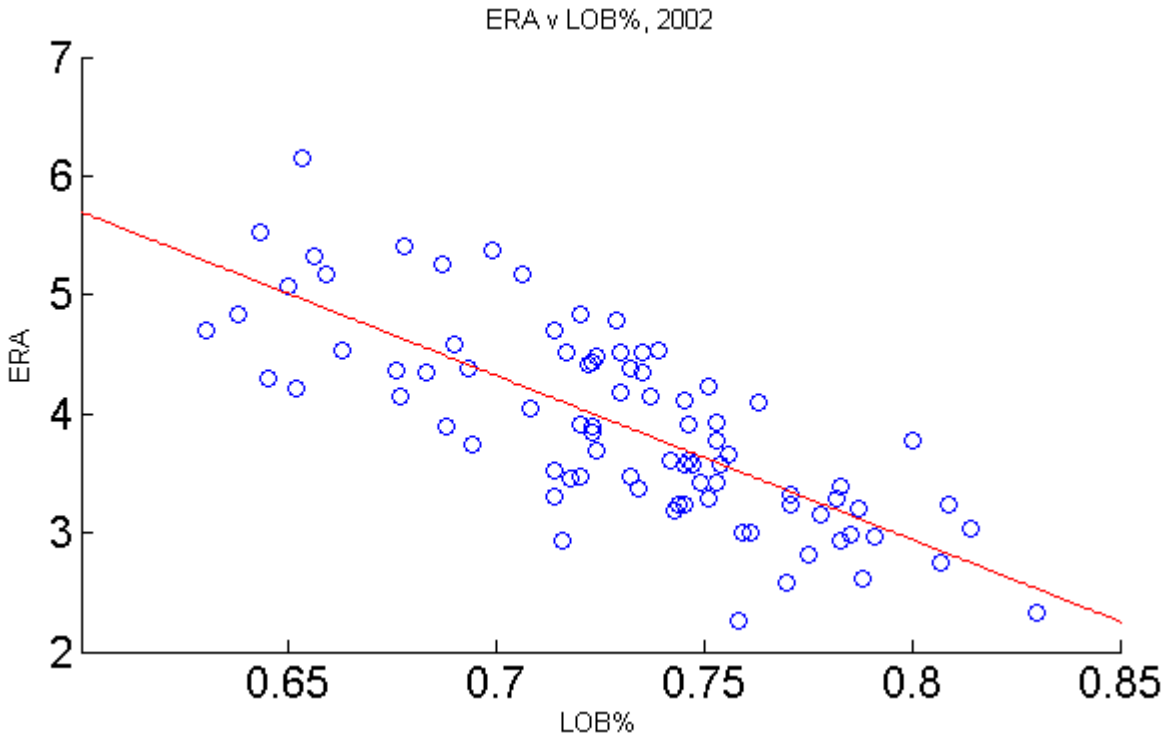

Figure 3.3: Scatterplot of ERA against left on base percentage, 2002

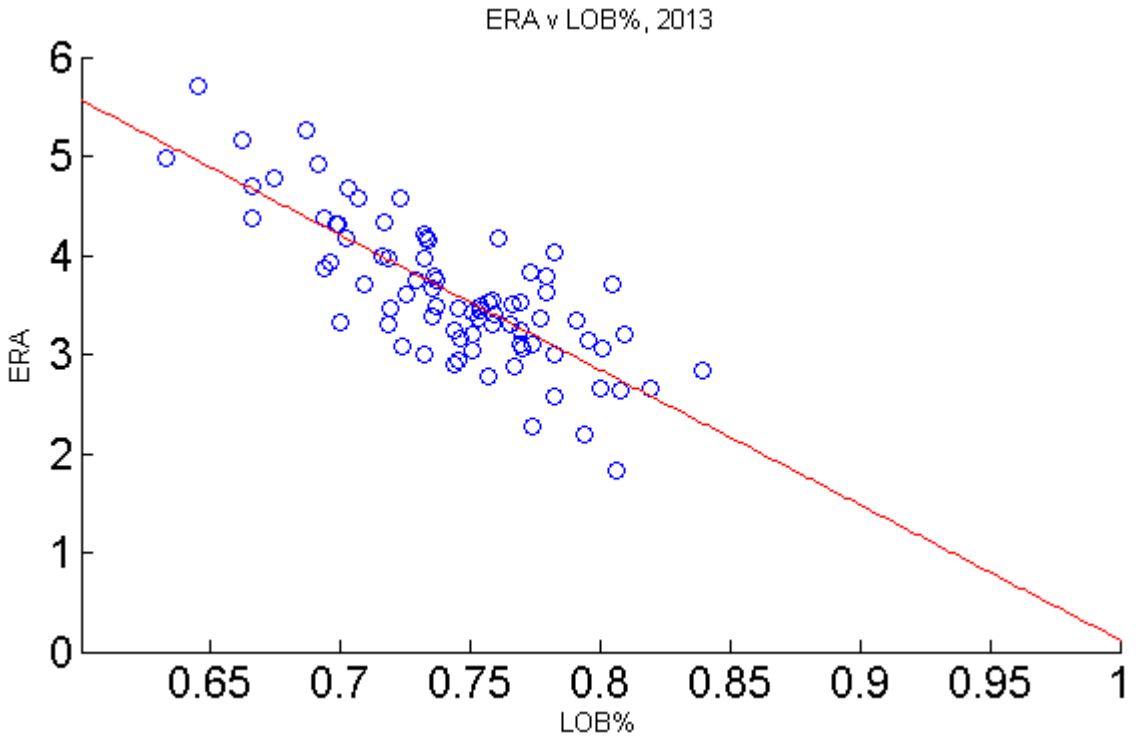

Figure 3.4: Scatterplot of ERA against left on base percentage, 2013

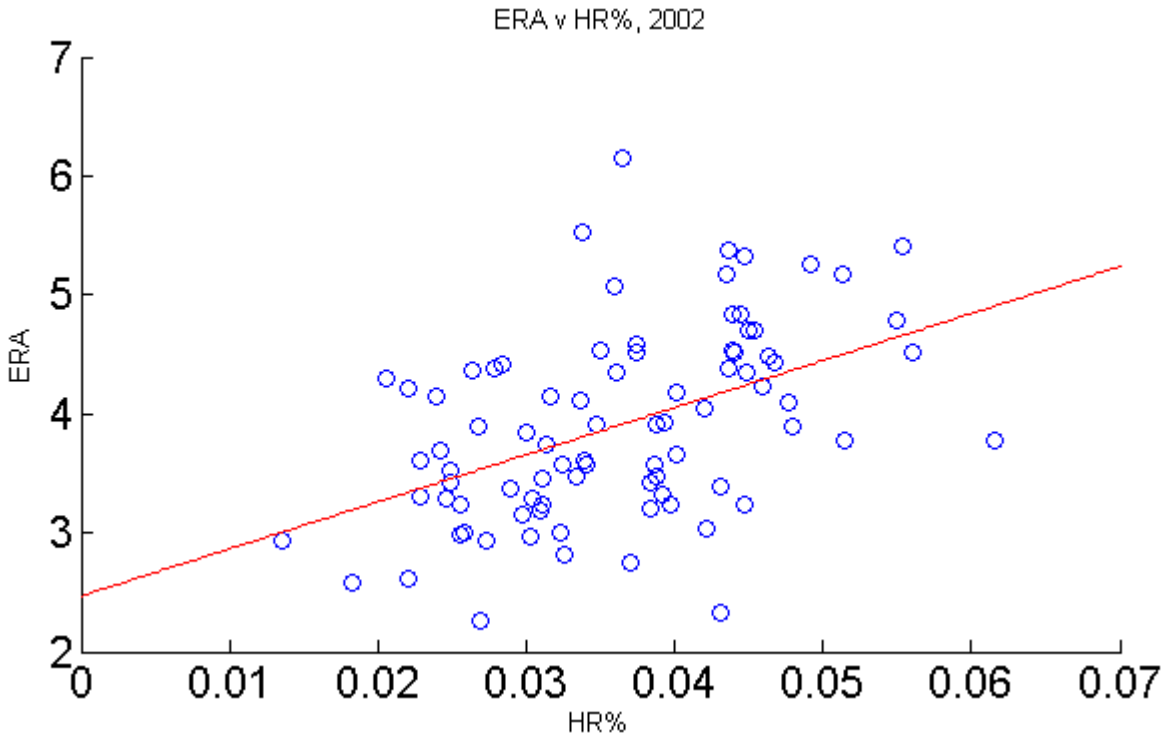

Figure 3.5: Scatterplot of ERA against home run percentage, 2002

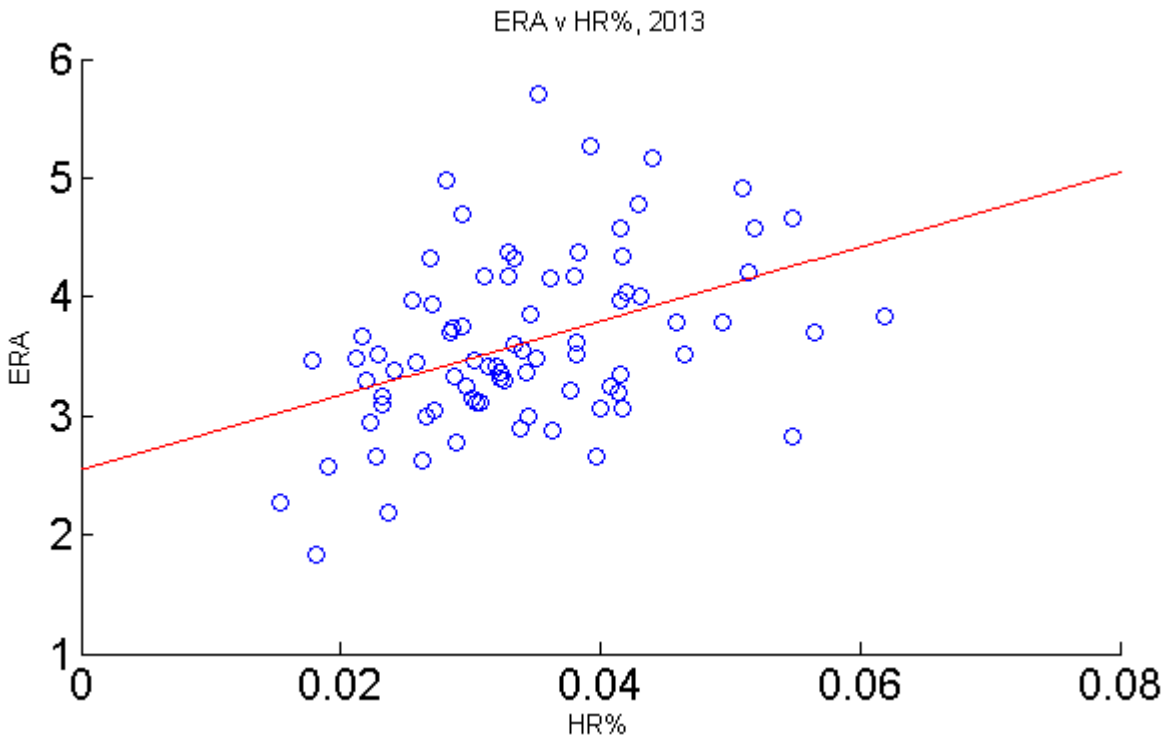

Figure 3.6: Scatterplot of ERA against home run percentage, 2013

Chapter 4

jFIP

4.1 New Data Inputs

My first proposal for improving the FIP model is by simply adding popups (PU) and hit by pitch (HBP) to the model. A strikeout is weighted in FIP as such because it is essentially an automatic out. The only way for a batter to reach base on a strikeout is through the catcher dropping strike three and the hitter beating the throw to first. Because it happens so rarely, this possibility is largely ignored. Pop ups (PU) are extremely similar. With very few exceptions, major leaguers catch pop ups. A ball gets lost in the sun every once in a while, but barring things like that, pop ups are just as good a source of easy outs as strikeouts. In addition, because they are by definition hit to the infield (balls in the air to the outfield are fly balls), it is nigh impossible for runners to advance on a pop up. Even if it is dropped, a fielder still has ample time to scoop the ball up and nab an advancing baserunner. The argument for including hit by pitch is much the same. A walk is weighted how it is because it is an automatic baserunner. There is no way for a batter to be put out when he has walked, at least not until he has safely reached first base. A hit by pitch is the same result, putting a batter on base without him having to do anything. A pop up is just as good as a strikeout and a hit by pitch is just as damaging as a walk, from the pitcher's point of view, so I propose to add these two statistics to the existing FIP model.

Because there are now that many more events that we are crediting to a pitcher, there must also be fewer events that we are not crediting to the pitcher. So by making the things a pitcher can control a larger part of the jFIP equation, we should be inherently tightening the model by lessening the effect that things like random variation can have on a pitcher's FIP.

To implement this, I am going to replace the strikeouts component of FIP with strikeouts plus pop ups and the walks component with walks plus hit by pitch. The coefficients of these variables will be left unchanged, at least at first, with any initial changes coming to the FIP constant. For each year in the study, league jFIP (before the constant) will be computed and used to calculate the new jFIP constant. Because the constant essentially represents the value of those events for which we do not assign the pitcher credit, we would expect that it would need to be changed, since we have just partly redefined which events a pitcher is and is not held responsible for. After calculating this new constant, jFIP will be computed for each pitcher and subtracted from his actual yearly ERA. These differences will then be squared and summed. This result will be compared to the result of subtracting current FIP from ERA and squaring and summing those differences, with the hope that jFIP's sum is lower.

Here is the current FIP equation:

$$\frac{(13 * HR + 3 * BB - 2 * K)}{IP} + C, \tag{4.1}$$

I am proposing this instead

$$jFIP = \frac{13 * HR + 3 * (BB + HBP) - 2 * (K + PU)}{IP} + C, \tag{4.2}$$

Because popups are almost certain outs and are a repeatable skill (i.e. pitchers show the ability to control their own popup rate, to an extent), we can remove them from balls in play and consider them equal to a strikeout, for which we know the coefficient is -2. The pitchers best at inducing popups still only record about half as many popups as the average pitcher records strikeouts, so the term should not be changed very drastically. The same thing is true of walks and hit by pitch. Certain pitchers are more prone to hitting batters than others, and the results are nearly indistinguishable from a walk. For this reason, I have lumped them in with walks and once again used the same coeffficient as before, 3.

4.2 Pitchers as Fielders

A second proposal for creating the new model is to incorporate pitcher defense in some way. Normally, we would not want defense in a statistic that has the words "Fielding Independent" in its acronym, but a pitcher's defense is ostensibly his own responsibility and, in theory, should be fairly correlated year-to-year. Defense is a skill for all baseball players, and a good fielder one year will most likely also be a good fielder the next.

But how to quantify defensive ability into a number that can be used in jFIP? Luckily, there exists a metric called Defensive Runs Saved (DRS). DRS sums a player's defensive contributions in different aspects of baseball to give a final defensive rating, generally a number from -20 to 20 runs, with positive values indicating runs "saved" or prevented above average, and negative values indicating the number of runs a player has cost his team. Since pitchers do not get nearly as many plays in the field as most other positions, the magnitude of pitchers' DRS is lessened, such that the range is cut approximately in half, from -10 to 10 runs.

Pitchers can accrue defensive value in a few different ways. They can make routine plays, such as fielding a comebacker to the mound, or covering first on a ball hit to the first baseman. They can help prevent stolen bases by holding runners on and having a quick delivery to the plate. Despite appearances, pitchers are generally more responsible than catchers for nabbing would be base-stealers; all major league quality catchers are extremely quick to catch, plant, and throw to second, thus the larger variability in pitchers' times to release the ball often determines if the thief is safe or not. Pitchers can also field bunts and simply not make errors to help save runs. For each of these 4 facets (making plays, not making errors, preventing stolen bases, and fielding bunts), a pitcher is credited for the runs he saves or costs his team. The runs are then summed to give a pitcher's cumulative DRS.

Once again, I visited www.fangraphs.com for the data. I created a custom leaderboard for each year with pitchers' innings pitched, DRS, and stolen bases. For some reason, the first year for which there exists DRS data is 2003, so 2002 will be omitted from the study for now. I had intended to work with stolen bases in some way, but since they are already incorporated into DRS there will be no need. Since DRS and its components are considered fielding statistics, they cannot be called in the same leaderboard as the pitching statistics imported into Excel before. In order to add a DRS column to

my existing spreadsheets, I imported the yearly DRS leaderboards into Excel and then alphabetically sorted the column listing pitchers' names. I also sorted the existing spreadsheets alphabetically by pitcher, and then simply copied and pasted the DRS column into the existing spreadsheet for each year, 2003-2013.

So we know that DRS will be incorporated into the jFIP model, but how? Sorting from highest DRS to lowest and doing a quick check of jFIP-ERA indicates that there is likely some sort of negative correlation between DRS and ERA, which is what we would expect. DRS is a measure of the runs a pitcher has prevented through his fielding ability, and fewer runs allowed necessarily leads to a lower ERA. One way to examine how DRS affects ERA is to take DRS at face value. It is not just an abstract number, but an estimate of runs.

Let's examine a hypothetical pitcher, who throws 200 innings and allows 100 earned runs. We calculate his ERA to be a pedestrian 4.50. Now imagine that instead of allowing 100 earned runs, this pitcher allows only 90. Now what is his ERA? It is 9*90/200 = 4.05, .45 runs better than before. Alternately, we could skip a step and just calculate the difference between his old and new ERA. It is simply 9*(100-90)/200. This is the effect on his ERA from allowing fewer runs. DRS tells us how many runs a pitcher saves or costs his team, and so we can use it to find the expected drop in ERA. The equation is

$$FRS = \frac{9 * DRS}{IP} \tag{4.3}$$

Fielding runs saved (FRS) is then subtracted from current jFIP to give a new jFIP.

$$jFIP2 = \frac{13 * HR + 3 * (BB + HBP) - 2 * (K + PU)}{IP} + C - FRS \tag{4.4}$$

4.3 Why IP can be misleading

Innings pitched is not the ideal denominator for a stat like FIP or jFIP. Baseball research focuses on very granular events, such as the outcome of one plate appearance. An inning is simply a collection of all those events before the fielding team records three outs. As such, all innings are not created equal; some will see only 3 batters and few enough pitches to track on one hand, while others can go on long enough that some batters come up twice before it's over!

The statistics that form jFIP are all counting stats. Every time a batter walks up to the plate, the pitcher has a chance to record a strikeout, or a walk, or a home run, or any other single plate appearance result. In addition, these counting stats are all completely separate, such that a pitcher can only record one per plate appearance. Therefore, the denominator we use to judge pitchers should be the total number of chances they have to record these counting stats. An inning can provide anywhere from three to 15 or more chances to strike a batter out, but a plate appearance provides one and only one. We must forget about how many innings a pitcher threw, and look instead to the number of batters he faced.

Total Batters Faced (TBF) is the pitching equivalent of batters' plate appearances (PA). Every batter-pitcher matchup that reaches a conclusion (the only way for this to NOT occur is if a runner makes the 3rd out on the bases or scores the game-winning run while the batter is up) counts as 1 TBF and 1 PA.

Because good pitchers allow fewer baserunners than bad pitchers, they also face fewer batters per inning. The effect of this on FIP is that the best pitchers are slightly underrated and the worst pitchers slightly overrated. To see why this is so, consider two pitchers who both throw 200 innings and strike

out 200 batters. Pitcher A is a terrific young hurler who only allows one baserunner per inning, on average. Pitcher B is nearing the end of his career and allows one and a third men to reach per inning. Who has better strikeout stuff? FIP sees them as having equally good strikeout ability, with both recording 1 strikeout per inning. But looking on a per batter basis changes the picture. Over the course of the season, Pitcher A faces 800 batters (3 outs per inning*200 innings plus one baserunner per inning), and pitcher B sees 867 batters. It is clear from this that Pitcher A is indeed the better strikeout pitcher, in fact there is about half a standard deviation difference in their strikeout rates when examining it this way, a not insignificant amount.

Pitcher A: 200 K/800 TBF = 25% K rate

Pitcher B: 200 K/867 TBF = 23% K rate

At first glance, there appears to be a very simple solution to this problem. First, compute the average number of batters faced per inning for all pitchers in the study. Then, calculate each pitcher's TBF/IP, compare it to the average rate, and adjust the pitcher's jFIP up or down accordingly. There are two reasons why this solution is not as preferable as it seems. The second one is minor and easily adjusted for (to be covered later), but the first is subtle and requires more thought.

The number of batters faced by a pitcher in any given year is biased by his fielders' success at turning batted balls into outs. If a certain pitcher had his fielders turn every single ball hit against him into an out, he would face far fewer total batters than if every single ball landed safely. In this way, TBF is influenced by something outside of a pitcher's control, which is completely against the point of FIP and jFIP. A pitcher does not have total control over his TBF, and so to calculate TBF/IP would be to intertwine a pitcher's skill with the ability of his fielders. He does have some control over TBF, though, and therein lies the answer.

Theoretically, a pitcher's goal is to get every batter he faces out. Whenever a pitcher strikes out a batter or induces a popup, he brings himself one out closer to the end of the inning. If he walks, hits, or allows a homer to the batter, he is no closer to reaching the end of the inning. If he allows the ball to be put into play, it may or may not provide him an out and bring the inning's end nearer. So we can credit a pitcher for keeping his TBF as low as possible by recording those automatic outs or by not allowing walks and home runs. Due to the effects of chance on balls in play, a pitcher may face many more batters one year than another, even if innings and underlying skills (strikeout rate, walk rate, etc.) are held constant.

What I did was to assume league-average success at preventing hits on balls in play for all pitchers. Given the rates at which they strikeout, walk, and hit batters (and allow popups and homers), I calculated the number of hits each pitcher would be expected to allow. I then compared this to the number of hits they actually allowed, and then added the difference to their actual TBF to create Expected Batters Faced (xTBF). (If a pitcher allowed fewer hits than I expected, his xTBF is higher than his TBF; if he allowed more then his xTBF will be lower.) Only then did I calculate xTBF/IP to find the number of batters a pitcher would be expected to face each inning, based only on his ability to record free outs and limit free passes. This way, the defense behind a pitcher has no bearing on his expected efficiency and jFIP remains fielding independent. Once xTBF/IP is known for all pitchers, the denominator of IP can be multiplied by each pitcher's xTBF/IP and then divided by the league average TBF/IP.

The second issue arises because of a quirk in the jFIP formula. Before the constant is added at the end of the equation, it is possible for jFIP to take a negative value, so we must be careful when applying a scalar. Clayton Kershaw faced (and was expected to face) far fewer batters per inning than average in 2013; his personal scalar is going to be less than 1. Thus, he will end up with a slightly less negative number before the constant and, after it is added back in, a higher jFIP, rather than a lower one. For this reason, I sorted pitchers by jFIP, lowest to highest. All those whose jFIP was less than

the jFIP constant were multiplied by the inverse of this scalar instead, so their jFIP would decrease as it should.

Chapter 5

Results

To test my proposal, I went back to the Fangraphs custom leaderboards and created one for each year of 2002-2013 with the statistics ERA, FIP, K, BB, HBP, HR, PU, and IP. I then took the total number of MLB K, BB, HBP, HR, PU, and IP for each year and plugged them into the jFIP equation in order to calculate the jFIP constant (by setting the equation to league-average ERA, as in FIP)(Equation 5.1). Once the constant was known, all pitchers' jFIPs were calculated (Equation 5.2) and compared to both their ERA and FIP. In equations 5.1 and 5.2, Lg in front of a statistic refers to the major league total for any given year.

$$C = LgERA - \frac{13 * LgHR + 3 * (LgBB + LgHBP) - 2 * (LgK + LgPU)}{LgIP} \tag{5.1}$$

$$jFIP = \frac{13 * HR + 3 * (BB + HBP) - 2 * (K + PU)}{IP} + C \tag{5.2}$$

The difference between FIP and ERA was computed for each pitcher and squared. The same was done for jFIP and ERA. The sum of the squared errors (SSE) was then computed for FIP and jFIP for each year (Equations 5.3 and 5.4), with a pitcher's actual ERA being the target value any deviations are measured from. The results are summarized in Table 5.2.

$$SSE_{FIP} = \sum_{i=1}^{n} (FIP - ERA)^2 \tag{5.3}$$

$$SSE_{jFIP} = \sum_{i=1}^{n} (jFIP - ERA)^2 \tag{5.4}$$

To see if the improvement of jFIP over FIP is statistically significant, we must perform a test of significance. In this case, a matched pairs t-test is the test of choice[2]. The sample size for each year is at least double the minimum n=40 and the data are approximately normally distributed, so the test is appropriate to use. For each pitcher for each year, we have his FIP error and his jFIP error. These will be our paired observations, and it is important to note that they are not independent. The null hypothesis would be that the mean of the FIP errors is equal to the mean of the jFIP errors. The alternative hypothesis is that the mean jFIP error is less than the mean FIP error, representing that jFIP is closer on average to ERA than FIP is.

To calculate the test statistic for this matched pairs test, I created a new column on the spreadsheet for each year. This column was simply a pitcher's FIP error minus his jFIP error. I then calculated

the mean and standard deviation of this column to use in the formula

$$t = \sum_{i=1}^{n} \bar{X} * \frac{\sqrt{n}}{s_x}, \tag{5.5}$$

where $\bar{X}$ is the mean of that new column, s_x is its standard deviation, and n is the number of qualified pitchers that year. These t statistics may also be found in Table 5.2. In each case, the column reading "t" is the value of t for the column preceding it. jFIP error refers to the jFIP formula with pop ups and hit by pitch, while jFIP 2 error refers to the new formula with defense added. These error terms were calculated using Equations 5.3 and 5.4.

Table 5.1: jFIP constants

Year	Constant
2002	3.23
2003	3.28
2004	3.31
2005	3.23
2006	3.39
2007	3.46
2008	3.35
2009	3.32
2010	3.29
2011	3.25
2012	3.29
2013	3.25

Table 5.2: t values

Year	FIP error	jFIP error	t	jFIP 2 error	t
2002	29.07	29.92	-.80		
2003	23.17	22.64	.64	23.59	-.26
2004	20.99	20.80	.23	17.25	2.16
2005	27.75	28.23	-.54	27.64	.07
2006	18.52	17.38	1.59	17.69	.72
2007	14.46	14.18	.48	12.64	1.53
2008	25.83	25.40	.59	28.17	-1.04
2009	25.17	24.49	.99	22.09	1.89
2010	26.96	26.48	.55	22.59	2.24
2011	26.24	24.42	1.82	22.08	2.38
2012	18.66	17.92	1.21	16.09	2.02
2013	19.18	18.33	1.34	15.62	2.17

jFIP has a lower total error than FIP in every year except for 2002 and 2005, interestingly enough the two years tied for the lowest jFIP constant. An easy way to test if this is a significant result is

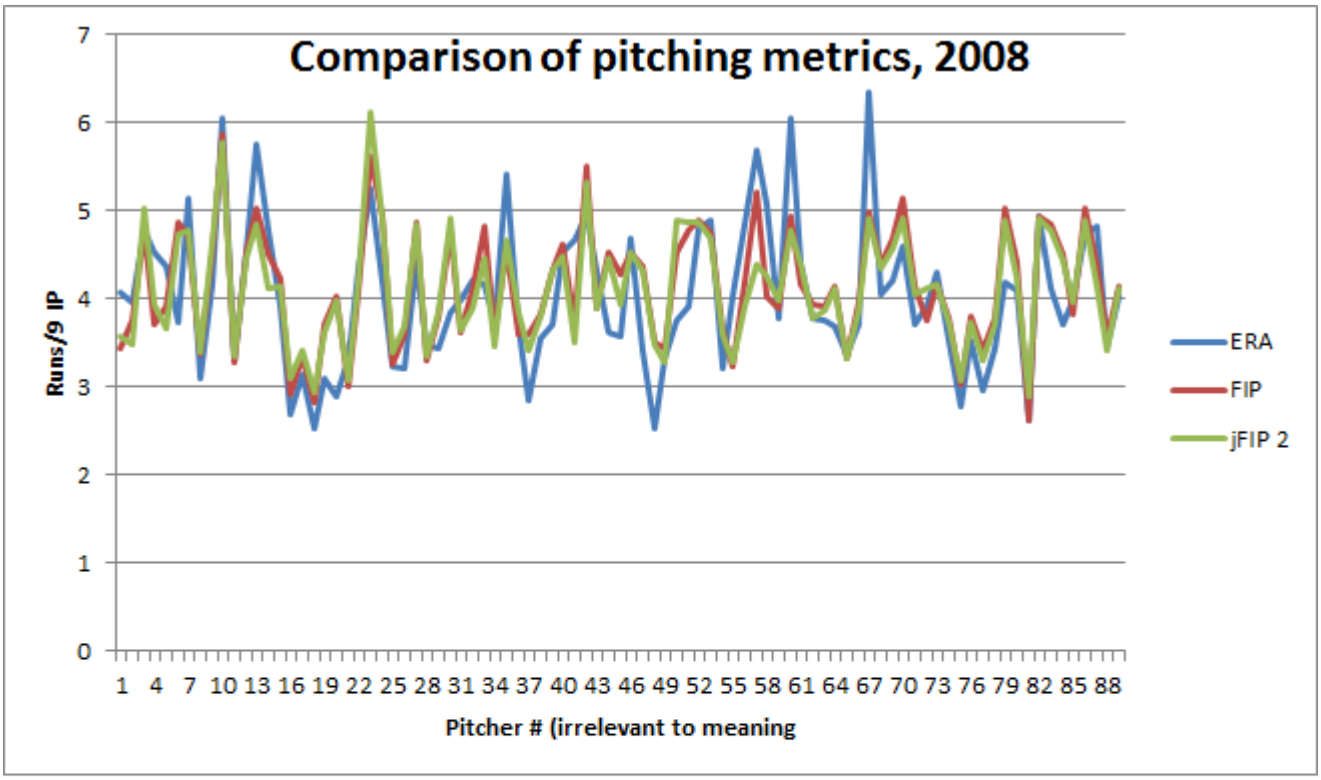

Figure 5.1: Trends in ERA, FIP, and jFIP 2 for 2008

through a binomial test. If jFIP were only as accurate as FIP, we would expect each formula to have a lower error in about half of the years, making it a binomial distribution with $p = .5$. The chances of seeing 2 years or fewer out of 12 with FIP having the lower error is just 2 %, so it is a reasonable assumption that jFIP is indeed better at getting closer to true ERA. Besides those two years, though, jFIP generally outperformed FIP by consistently having 2-3 % less total error. These results are consistent with what I had expected; there was a slight increase in the accuracy of the formula with the inclusion of pop ups and hit by pitch. These events are rarer than strikeouts and walks, respectively, but have much of the same value to a team in a game. As such, jFIP is more accurate for more pitchers, because more of their own results are now being solely attributed to them. At the same time, though, this is only a small increase because of the relative rarity of pop ups and hit by pitch.

Figures 5.1 and 5.2 show trends in the three pitching metrics. The line associated with ERA has the most spikes, both positive and negative, because of all the random deviation present in ERA. The FIP line is much smoother than for ERA, showcasing its reputation for being more stable year-to-year than ERA. jFIP is smoother than ERA, like FIP, but also contains hints of the spiky highs and lows of ERA, thereby explaining some of the difference between a pitcher's FIP and ERA.

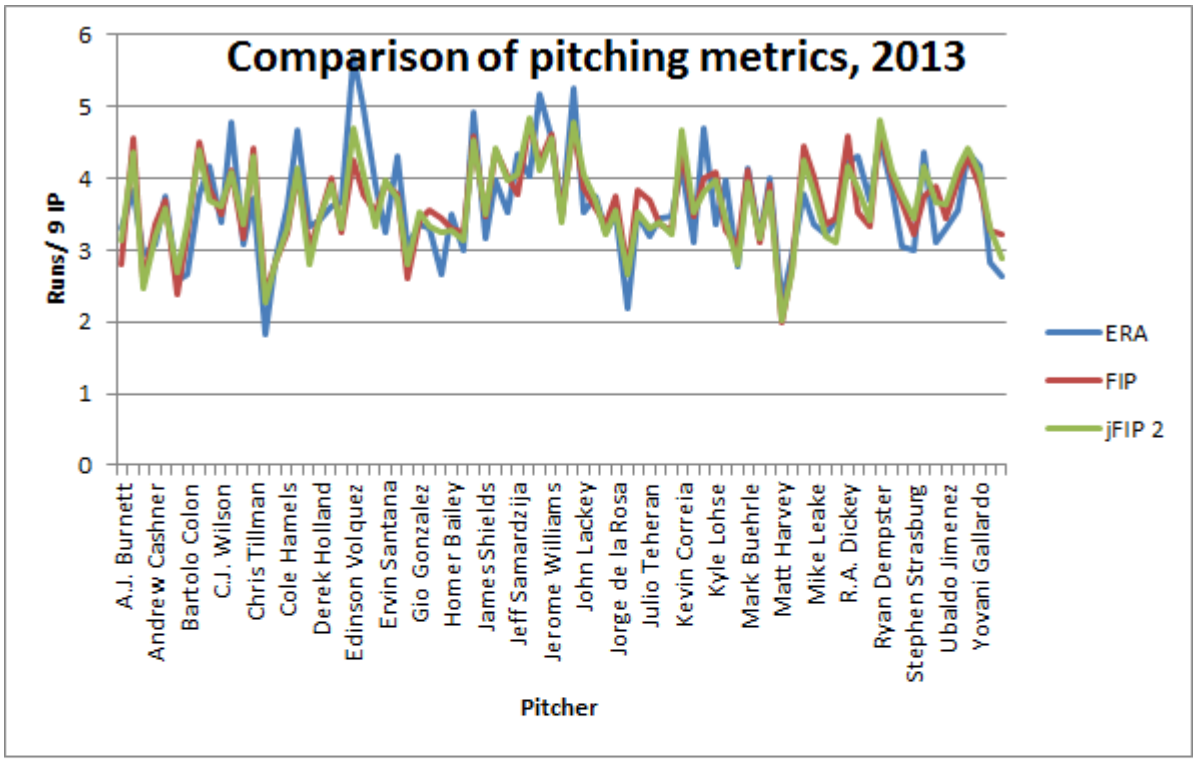

Figure 5.2: Trends in ERA, FIP, and jFIP 2 for 2013

Chapter 6

Discussion

Using a table of critical values for t, we see that the rows for degrees of freedom jump from 60 to 120. Since the number of qualified pitchers lies in this range every year, we pick the lower threshold, 60, since that will make it more difficult to surpass the critical values. The critical values are not very different between the two rows anyway. Choosing a 5% significance level, we see that the critical value is 2.

For the initial jFIP proposal, none of the years in the study show a statistically significant improvement, with two years actually representing jFIP being worse. But out of the 11 years for which we can calculate jFIP 2, which incorporates fielding, 5 of them have t statistics that exceed this critical value, with another very close. The previously discussed scalar for innings pitched did not have any meaningful effect on the results. The difference in batters faced per inning (the best of the best only faced about 5% fewer) may just be too smal to make a difference. So for about half of the years in this study, jFIP 2 is better than FIP at predicting ERA in a statistically significant way. If only one had managed to surpass the critical value, it could potentially be chalked up to random chance, but five years shows that there is something meaningful here.

Due to the large coefficient for home runs and their relative rarity compared to the other components of jFIP, a sensitivity analysis was performed to ensure the model is robust enough to withstand changes in MLB's run environment and still be a useful tool. For each year, 2003-2013, I re-did all the calculations involved in finding the jFIP constant, C, while changing the coefficient of home runs to 12 and then to 14. With the coefficient at 12, the constant increases .11 or .12 per year, and it sees an equal decline when the coefficient is set to 14. In both cases, though, the relevant t statistic only changed by a couple hundredths and there was no change in the significance of the results. The jFIP equation appears to be robust enough to be used in future years, barring an unprecedented change in the run-scoring levels of MLB.

Bibliography

[1] *Fangraphs*. 2014. URL: http://www.fangraphs.com.

[2] I Miller and M Miller. *John E Freund's Mathematical Statistics*. Boston, MA: Simon & Schuster, 1999.

[3] P Palmer and J Thorn. *The Hidden Game of Baseball*. Boston, MA: Doubleday, 1985.

[4] T Tango, M Lichtman, and A Dolphin. *The Book*. Boston, MA: Potomac Books Inc., 2007.

[5] Tom Tango. *Deconstructing FIP*. July 2011. URL: http://www.insidethebook.com/ee/index.php/site/comments/tangos_lab_deconstructing_fip/.