Ben Zeman

STATS 406

Prof. Fredrickson

04-24-2021

How To Hit the Jackpot with a Pitching Prospect

Background and Motivation

The best baseball players in the world are notable for four aspects of their game: raw talent, situational knowledge of the game, perfected skills, and work ethic. You may hear about a player in the news who lacks one of these qualities—in comparison to his colleagues, I will strongly emphasize—but the reason these players make the news is because it is rare! With a few exceptions, you need to have all four of these traits to be a successful player.

Now, you may be thinking that I am going to analyze all four of these player traits, but I will only focus on one: raw talent. For those unfamiliar, "raw talent" references skills that can be acquired relatively easily with the right athletic build. For instance, an NFL linebacker may only have to practice his swing for a few hours, and he could probably hit a home run, while your average man could practice swinging for weeks and still never hit one. The linebacker has not worked harder, he has more athletic gifts. Interestingly, while a manager of a major league team generally focuses equally on the four traits, a scout who analyzes young process will focus heavily on raw talent. This is from the philosophy that the other three traits are teachable but raw talent is not. For pitchers, raw talent basically boils down to pitch velocity. Sure, lifting weights can help a little, but the average man can't throw 90 MPH no matter how hard he trains. Any

player can be taught accuracy and off-speed pitches, so those qualities aren't as important to scouts. Therefore, this study focuses solely on pitch speed.

In this paper, I create and analyze a method of predicting the speed of fastballs—the pitch where speed matters most—based on a player's height and what year he is pitching in. I am not explicitly given height data, so I estimate height as a function of other variables, then I estimate pitch speed as a function of the height function and time. With the height function, depending on the variances of my estimates, I can predict how hard a player can throw, which is invaluable for scouts who have never seen a pitcher play and want to choose who to watch. This model could also help analyze players who are recovering from an injury and have not reached their fastball potential.

The time variable is useful in all cases, from an early high school prospect to a young major leaguer. Scouts can't just analyze how hard he pitches now; they need to predict how hard he will pitch if and when he is a player or even an ace on an MLB team. Now, the younger the player, the more extrapolation is involved, so scouts should proceed with caution when using this model for very young prospects. Therefore, these models must be made carefully, and it imperative to not "force" a time dependence when the correlation is negligibly week. However, if you understand the variability that comes with this model, I believe it can be a very effective tool.

Data

Year	Vertical Pitch Location (FT)	Vertical Angle (°)	Release Point (FT)	Velocity (MPH)
2015	2.963	34.685	2.970689395	84.1
2015	2.347	34.225	2.354558193	84.1
2015	3.284	35.276	3.29186012	85.2
2015	1.221	28.354	1.226996232	84
2015	2.397	32.274	2.404017098	84.8
2015	1.61	31.469	1.616800631	85.3

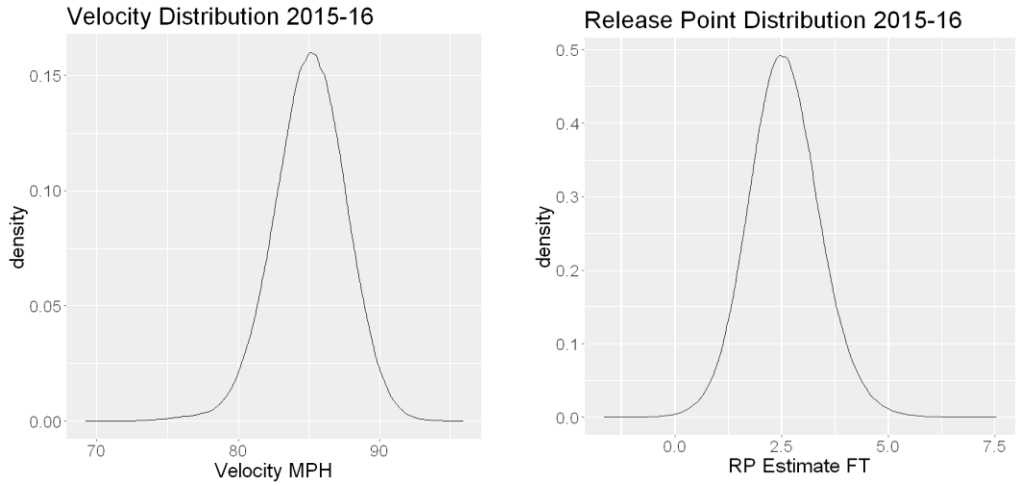
Figure 1: Preview of the subset of our data frame used in this analysis

Year	Vertical Pitch Location (FT)	Vertical Angle (°)	Release Point (FT)	Velocity (MPH)
2015	2.565393774	29.8872711	2.571810717	85.06181178
2016	2.569961463	29.48648899	2.576275901	85.06320584

Figure 2: Average Values of our Statistics by Year

I would be lying if I were to say the data set I used was exactly what I was looking for. It did have a large number of properties, but most were irrelevant to this study. The only reliable source of pitch-by-pitch data was from Kaggle, where it was transferred from MLB.com. The data was described to be all pitches from 2015-2018, but 2017 and 2018 were omitted from the dataset. Each row in the set represents a pitch, and the properties that were useful for this study were the speed of the pitch when it crossed home plate, the height of the pitch from the ground, the year that the pitch took place in, and the vertical drop angle of the pitch when it crosses home plate. Only the pitch final velocity was directly used; the other variables were used to create new properties which I used directly in my analysis. As you can see below, the two main distributions we are focus on are approximately normal, allowing us to utilize many assumptions we have

used in the course. They are slightly different distributions, which comes across in the following chart, but they can be simulated in similar ways.



Figures 3 and 4: Velocity and Release Point Distributions

I was unsatisfied with the raw data itself, so I performed a bit of manipulation. As I stated before, I was only analyzing fastballs, since they are what demonstrate how hard a pitcher can throw. The dataset contained all pitch types, so I filtered on fastballs. Second, I wanted an independent variable representing, or even correlated with, height. This was not available. However, the pitcher's release point—the height from the ground that the ball leaves his hand—was available indirectly. These are different statistics, but my thought process behind the correlation was that a pitcher will release the ball at a spot relative to his shoulder or chest, not the average pitcher's shoulder or chest. Therefore, with shoulder height dependent on height, release point would naturally be dependent on height. However, as an aspiring statistician, I had to take this with a grain of salt. Anyways, I estimated the release point for each pitch with basic trigonometry:

$$RP = y + \tan(break_y)/90$$

The units are feet. Y is the vertical coordinate, break is the angle to the horizontal that the pitch crosses home plate at, and the pitcher's mound is 90 feet from home plate. Now, the pitches weren't explicitly marked by year, they were marked by at bat ID, which I could easily extract the year from. At this point, with only end velocity and the columns I created, my data is clean and ready for analysis. If you remember, our main goal is to find velocity as a multi-faceted function of release point, so we start by trying to observe the relationship between the two.

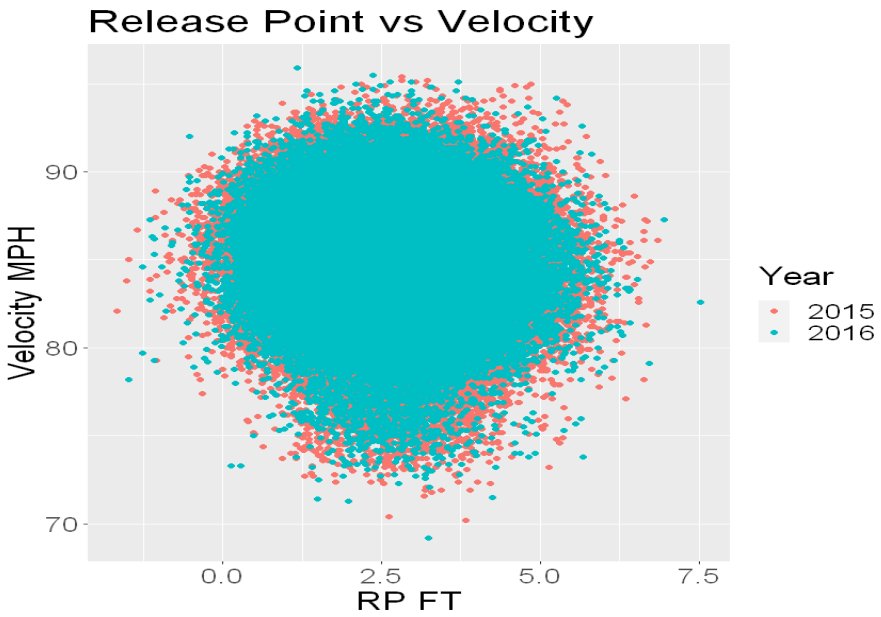


Figure 5: Velocity vs Release Point

As you can see here, it is tough to distinguish data points due to the sheer volume of data, which makes it hard to spot any correlations. I can observe subtle distribution differences, but I need to perform an in-depth analysis to create a meaningful model.

Methods

After cleaning up the data, my first goal was to estimate the height of the player throwing the given pitch. I quickly realized I didn't have enough data at the moment to predict a constant value for the height, but I knew I was going to use multivariable optimization in the future, so I

was content with leaving one variable unsolved for at the moment. This variable would be set to optimize a different value. Now, even though I had an unknown variable to store, I still did not have height data, only release point estimates. However, I could use permutation methods to calculate an explicit function relating height (H) and RP (Y). Since this function had to be explicit and the H/Y dependence was unknown, the function had to be, more generally, an explicit function of Y and a function of H. To make H a function of a product, I created:

$$T = \frac{f(H)}{Y}, f(H) = YT\ , H = f^{-1}(YT)$$

I describe my analysis of the T vector and the functional form in my analysis. With this function, I do a permuted cross-validation predicting H to prepare for future optimization, since it returns predicted values and error terms. Using this predicted H value, I do another CV to prepare for optimization, estimate velocity (A), and return prediction errors. For this CV, A is a function of H and time. Here is the equation:

$$A = g\left(f^{-1}(YT)\right) + kt = g(H) + kt$$

For the second CV, I needed to add in a time component. Since I had time data and pitch speed data, I could do a difference of means bootstrapping method. The time component in the above function would then be linear with a slope of the bootstrap result.

The final algorithm was what tied it all together. The two unknown functions above each had an unknown parameter. Therefore, I iterated through some parameter values and chose the values which corresponded to the minimum combined error from the H and A linear models. With these values we could evaluate or models for H and A respectively, measure descriptive statistics, and interpret my findings.

Simulations

My first simulation process was to find a distribution for the dummy vector T. My data was set up perfectly for distribution-free permutation, since the function of H is unknown. I had to create a dummy vector, Z, with binary data which shows whether the value is coming from the H or Y distribution. Z is a permutation invariant, since the H-Y combined sample does not have indices which are related to each other in any way, which allows us to do this permutation. I used random ranks for the unknown function and calculated ranks for Y, combined the sets, and used the standard permutation method with the quotient mean of the two randomly selected RV values as the test statistic. I describe this further in the Analysis section.

Since YT is a variable of interest but they are different sizes, I take a sample of T with replacement that is the size of Y and perform bitwise multiplication. This is okay because statistics from T are not relevant, only the distribution, which we can freely sample from. Now, I think $H = f^{-1}(YT)$. I want to find this function, specifically, a function with minimal error. Now, I can't forget optimization of pitch speed, so I will leave the function in general form for now. With all of the missing data and estimations, I want to limit assumptions as much as I can, unless the assumptions are clearly true. I'm not very sure about the relationship between H and YT, since T is not an observable, merely a created statistic. I concluded that an nth order polynomial was reasonable since it is the most flexible of relationships to my knowledge. This is assuming the relationship is in valid functional form, but I believe that is reasonable. I implemented an nth order polynomial function of a vector and n, leaving n to be changed later. Now, to help with future optimization, I wanted to estimate the error of predicting H as a function of n. I used cross-validation for this, as it is a very effective way to create a vector of squared prediction errors, a vector of predictions, and a vector of model coefficients for the

functions. This algorithm assumes dependence which I have already assumed, but it doesn't assume an error distribution, which is important with all of my estimation. I chopped my data in half, into a train and test set, trained a linear model of the initial value of the H function vs the permuted value, made predictions with the test set, and returned the three vectors mentioned above.

For A, I did another cross-validation, where I used a similar process but my linear model was now pitch speed as a function of $\left(f^{-1}(YT)\right) + t$ with general a. This linear model is more telling than our last one because we have real values for our dependent variable, not just estimates. If you use this model for prediction, make sure to keep this in mind. We return the same three vectors.

Our third simulation, the final process before optimization, is estimating the time component of A. I assumed that since A is approximately normal, both subsets-by-year are also normal-like. Therefore, the means and medians will be representative of the distribution. This is a relatively safe assumption considering many other distributions also are well represented by these statistics. Of course, some aren't, and maybe that is something to experiment with further. On one hand, I wish I had more than two years of data to work with. On the other hand, the binary nature of the year column allows us to predict the difference in means for pitch speed using a two sample bootstrap. The number of pitches for the two years are unequal, so I had to sample from the larger sample without replacement to make the two samples equal in size. I performed a two sample bootstrap for the samples from the two years, compared the mean and median confidence intervals, and then used a sample from the smaller interval as a vector for the linear model of A.

Analysis

To start off my analysis, I need to discuss the T statistic created from permutation, as it is used throughout the study. The permutation function returned a distribution for T which was smaller than the other distributions. This function is basically a dummy function as its' characteristics and distribution mean nothing on their own, only when the distribution is coupled with Y to form the independent variable in the definition of H. At this point, I can ignore the function of H that was needed for permutation, all we care about now is the inverse function defined above, as we can think of H as a function of YT. We are assuming dependence because the inverse function naturally has dependence as it was the product of the permutations.

For my first CV which helps find H, I would have liked to have a more trustworthy model with actual H values, but I had to work with the table I was given. I can't be too confident in these findings, but they build up to the major research question and provoke ideas for future studies.

Now, for the second CV in relation to A, I assumed that unlike for f^{-1}, g would be an increasing and concave down curve. This comes from my knowledge of baseball and indirectly from physics: we know that taller people throw harder just from observation (increasing) but we also know that while pitchers in their teens can jump from 60-90 MPH relatively quickly, there appears to be a large difference in ability when comparing 95-100 MPH. Therefore, it is reasonable to believe that factors like height have a greater effect on pitch speed the lower the initial speed is (concave down). With any assumption, however, we can always question what the results of the study would be without the assumption, especially if the results show weak or no correlations. A common function with these properties is a logarithm with the argument multiplied by a constant a. All log functions fit these properties, so the base generalization helps optimize the function.

After these three simulations, I had two unknown functions. These unknowns were both stopping me from simulating and not allowing me to predict H or A. Therefore, optimization was necessary. The only unknowns were n and a, and I wanted to limit prediction error. To limit prediction error for H and A simultaneously, I iterated over reasonable n-a combinations, calculated the sum of the ranks of the mean values for H and A error, and chose the n-a combination with the highest combined rank. n must be a positive integer, but I used positive integer values of a to limit computational cost and for my own convenience. With a more powerful computer and more time on my hands, I could experiment with a larger range of a values. With n and a, I solved for predicted H and A, the coefficients of both linear models, and the errors in both linear models. The errors can be used to justify when this model should be used and when it is inappropriate.

However, before even running the optimization algorithm, I made a conclusion that significantly altered the model, and in fact made the model significantly simpler. I thought that I was going to see a significant difference in pitch velocity from year to year, as we are seeing more 100 MPH pitchers than ever these days. However, this was not the case. After 1000 bootstrap replications—any more would be too computationally expensive and the CLT likely holds here—I found that the value of 0 was not just in, but right in the middle of, my confidence intervals for differences of mean and differences of median from data stratified by year. Specifically, the CI for mean difference was (-0.0142, 0.0268) while for median it was (-0.0389, 0.0497). These CIs were at 95% confidence in units of MPH. I didn't want to make the alpha level any lower, as there is already likely to be significant extrapolation error when this model was put to use. I did not want to try to "force" a time dependence; I would rather make my model independent of time. So, I adjusted my linear model:

$$A = g(H)$$

While this removal may not help for calculations and quantitative comparisons, it may help for qualitative comparisons by warning analysts to not expect a raise in pitch speed without first proving my results to be inaccurate. A poor model is often worse than no model, or what many call, the "eye test".

For my optimization algorithm, the n and a values that minimized the balanced squared errors of H and A were both 2. First, I constructed the formula for H for n = 2 value. The polynomial regression is shown here:

$$H = 183.14 - 0.0002(Y^2T^2 + YT + 1)$$

I was not able to figure out a method of optimization which changes function parameters and uses the linear model function to find parameters for each term. Maybe, if I could do so, I could find more of a correlation between H and YT. With an estimated value of 183.1 feet for H with a tiny variance of 0.0009, it doesn't appear that our permutation optimization tactics helped us discover that height changes significantly with our created variables. In addition, the mean height is unrealistic, which is likely the result of not having height data and relying almost purely on assumptions. That doesn't even include the astronomical MSE of 20349 feet. There was a major risk of a result like this happening, and while they sometimes pan out, this one didn't. From this, there is no way we can conclude any correlation, as changes in H would nowhere near approach the root of the MSE. Showing a plot here would be counterproductive, as I wouldn't want anyone to interpret much at all from this model. Despite the unrealistic values, we hope to still find a relationship between H and A. The logarithmic regression result is shown below:

$$A = 84.587 + 0.079 * \ln(2H), A = 84.587 - 0.079 * \ln(366.28 - 0.0004(Y^2T^2 + YT + 1))$$

This is not pretty by any means, however, the y intercept and mean prediction are around where your standard fastball would clock in at, with the prediction at 85.05 MPH with a very small variance of 3.21*10^-8! Now, this could be a result of similar data values, so I still have to verify that I performed significant analysis. The MSE of 6.68 pales in comparison to that of H, even scaled to units. Now, to see if this model is realistic, we can imagine a hypothetical scenario.

A player is 6 feet tall. He would be projected to throw 84.78 MPH. This makes a lot of sense, as it is higher than the minimum, but less than average, since 6 feet is on the shorter side for pitchers. Now a 6'6 pitcher, who is more common in the MLB, would throw 84.79 MPH. This is a tiny increase, nowhere near the root of the MSE, which is a measure of determining significant change. A 5'6 and 7'0 pitcher, both of which have nearly but not quite happened, only have a 0.01 MPH difference.

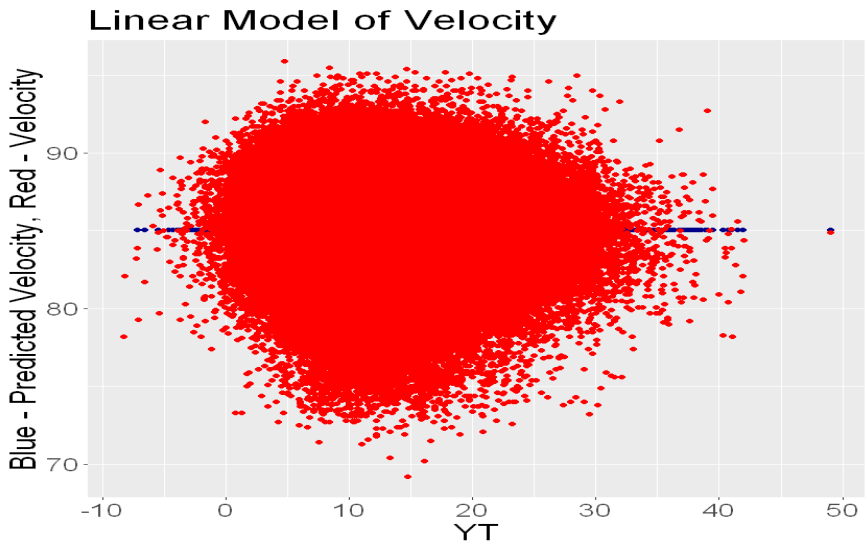


Figure 6: Linear Model of velocity as a function of YT

As we can see from this graph and hypotheticals, it would be insane of me claim a correlation between height and pitch velocity. Therefore, even without a strict confidence interval, I know I will not reject the null hypothesis that height and speed are independent.

Discussion

My research did not provide you with any models that I would suggest using for quantitative calculation. I did not disprove any null hypotheses, in fact, I can't come to a conclusion. The height-release point regression model did not confirm much other than the fact that it is very tough to replace missing data. Attributes that you believe might correlate don't often give you the conclusive results you want. However, some statistics did surprise me in a positive way. A positive slope coefficient for height, despite the fact that I did not have height data, could hint at a correlation which could be uncovered in future studies.

This study was certainly not without its flaws. Some could have been avoided easily, some more difficult. The first was data quality. I didn't just pick the first dataset I saw; I think this was the only publicly accessible pitch-by-pitch data with complete entries. Nevertheless, I was able to create multiple models with it. The data quality lead to estimation of an independent variable H, which was bound to cause uncertainty in my final model. Another major source of error was minimal repetition. If you can see my code, my permutation only has 80 replications and the CVs are only done once after permutation given an a-n combination. The bootstrap for time dependence was well covered, yet ironically was the least conclusive. I originally planned on doing repetitions of the CVs, however, they proved extremely computationally expensive. Either it would take several hours to run, for which testing is near impossible, or it would fail for lack of memory. My computer may have been a major issue, but it was my only option since this project requires software on my hard drive. The one benefit, however, of having these glaringly obvious improvements I could make, is that the research can easily be carried on by myself or others who are inspired by it. Some improvements I could have made were saving variables to

permanent memory so they wouldn’t be lost with re-runs and finding a way to optimize my laptop speed before starting this project. Some future projects I could consider would be to finesse non-public data and do a time analysis, automate the program to run in the background with more repetitions, or try fitting models other than logarithms and polynomials. Some parts of this project went better than others, but they can all lead to future knowledge. We might not yet know how height effects velocity quantitatively, but we can use our qualitative instincts to keep trying to show a relationship. Until then, we have to settle for this.

Works Cited

Fredrickson, M. F. (2019). *Permutation Methods*. Lecture Notes STATS 406. https://umich.instructure.com/courses/417271/files/folder/Lecture%20Notes/Week%2009?preview=18446940

MLB Pitch Data 2015–2018. (2020, May 17). Kaggle. https://www.kaggle.com/pschale/mlb-pitch-data-20152018#pitches.csv

Pitcher | Baseball Positional Guidelines | Go Big Recruiting. (2016). Go Big Recruiting. https://www.gobigrecruiting.com/recruiting101/baseball/positional_guidelines/pitcher#:%7E:text=Prototypical%20Division%20I%20pitching%20recruits,it%20every%20once%20and%20awhile.