Predicting a Major League Baseball Pitch

April 15, 2020

Abstract

For a Major League batter, one of the most challenging aspects of the game is recognizing the pitch type. As baseball becomes more technologically advanced, the need for analytics is now much more crucial. This study evaluates the relationship between the type of pitch about to be thrown and factors such as the location and velocity of the previous two pitches. Using three pitchers, Hyun-Jin Ryu, Gerrit Cole, and Michael Wacha, this study creates a logistic regression model to observe the relationship between these factors at an individual level. The results of this study provide insight into a pitcher's most telling attributes, also allowing us to compare the significant trends between multiple pitchers.

Introduction

For a Major League Baseball (MLB) pitcher, the velocity and break of their pitches play a monumental part in determining success. However, attempting to predict success using only these factors can often be lacking. Hitting a baseball is a skill that relies heavily on reaction time. As research has shown, the sensorimotor abilities of a batter have significance when predicting statistics such as on-base percentage, walk rate, and strikeout rate (Burris et. al., 2018).

This information seems obvious considering that a 90 mile per hour fastball gets to the plate in around 0.4 seconds. However, for an MLB batter, proper timing makes hitting this easily achievable. The purpose of an off-speed pitch is to challenge the hitter's timing through movement and deception. Considering that pitchers often fall into tendencies when selecting a pitch, the ability to predict pitches would give a batter a significant advantage while at bat.

This study intends to ask about the relationship between the type of pitch about to be thrown and factors such as the location and velocity of the previous two pitches. Possible impacts include the improvement of pre-game scouting reports and player evaluation by general managers and fans.

Methods

Participants

This paper will observe the 2019 season of three selected MLB pitchers (Hyun-Jin Ryu, Gerrit Cole, and Michael Wacha). To create an accurate sample of the 2019 season, we will take every pitch thrown from 8 randomly selected starts to train and test the logistic regression model. We are using a sample of 8 because the data was often presented out of order. Each start had to be manually sorted and checked for accuracy. Since our analysis requires at least two prior pitches to be thrown, we can only use pitches after the second pitch in the at-bat.

Data Collection

Data will be obtained from PITCHf/x using the download service provided by Brooks Baseball (www.BrooksBaseball.net). The data gathered by Brooks Baseball is initially downloaded from MLB Advanced Media (MLBAM) and contains a set of pitch classifications obtained through the use of neural networks. Brooks Baseball then manually reviews each pitch for accuracy by checking with additional sources, such as video replay and on-field personnel.

The dataset is rather clean with very few missing values. Since there is such a large amount of data, observations with crucial missing values will be excluded from our analysis.

Variables/Description of Analysis

With the variables in the equation defined in Appendix Table 2 we ran the following logistic regression model to determine the probability of a given pitch being off-speed:

$$\begin{aligned} logit(Y_i) = {} & \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 X_1 + \beta_5 X_2 + \beta_6 X_3 + \beta_7 X_4 + \beta_8 X_5 + \beta_9 X_6 + \beta_{10} X_7 + \beta_{11} X_8 \\ & + \beta_{12} X_9 + \beta_{13} X_{10} + \beta_{14} Z_1 Z_2 + \beta_{15} Z_1 Z_3 + \beta_{16} Z_1 X_1 + \beta_{17} Z_1 X_2 + \beta_{18} Z_1 X_3 + \beta_{19} Z_1 X_4 \\ & + \beta_{20} Z_1 X_5 + \beta_{21} Z_1 X_6 + \beta_{22} Z_1 X_7 + \beta_{23} Z_1 X_8 + \beta_{24} Z_1 X_9 + \beta_{25} Z_1 X_{10} + \beta_{26} Z_2 Z_3 \\ & + \beta_{27} Z_1 Z_2 Z_3 \end{aligned}$$

Because pitchers often factor the handedness of the batter into their pitch selection, it is important to monitor its relationship to the other variables. As such, we have included interaction terms between handedness and all other predictors.

An important thing to note is the lack of previous pitch type and the addition of previous pitch velocity. This is solely to provide more information to the model. Many pitchers have a large range of velocity between their off-speed pitch selection. The addition of velocity to the model allows it to utilize more specific information.

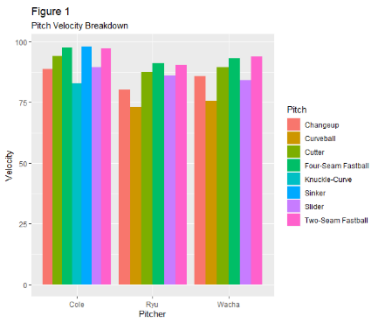
We used the statistical software Minitab to select our predictors and R 3.6.1. to obtain predictive accuracies and analyze the models. To obtain our regression coefficients, we used a 10-fold cross-validation to minimize error from picking a nonrepresentative training dataset. To select predictors, we utilized forward selection based upon Akaike information criterion (AIC).

Results

Description of Variables

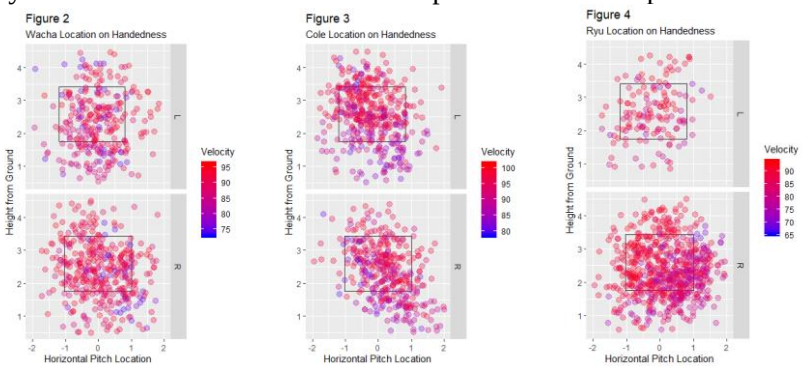
With our response variable (pitch type) being based on the MLB classification of the pitch, pitches were grouped as shown in Appendix Table 1.

Based on Appendix Table 1, all three pitchers have a rather similar pitch mix. To be able to differentiate a little further, we can look at the velocity of each individual pitch category. The information is broken down in figure 1.



For the purpose of our model, horizontal and vertical locations are measured separately. Rather than only using the location as measured from the center of the plate, horizontal location is split into two variables: the distance from the center of the plate to the batter and the distance from the center of the plate away from the batter. Vertical distance is measured in a similar fashion. It is determined by the distance above and below the center of the batter's individual strike zone as provided by Brooks Baseball.

We can see that even though Cole has a similar pitch repertoire, the average velocity for each of his pitches is higher than the others. Wacha and Ryu, in addition to having the same pitch repertoire, have very similar velocity distributions. A breakdown of the pitches in our sample is below in Figures 2-4:



These figures show the location of every pitch in our sample according to batter handedness. Interestingly, the number of pitches that Ryu threw to left-handers appears rather sparse. This can be explained by the phenomenon that a batter of the same handedness as the pitcher is unable to see the ball as well as a batter of the opposite handedness. Since right-handed batters are much more common, this effect (as shown by Figure 2 and Figure 3) becomes more negligible with right-handed pitchers. The mean location for each pitcher is shown by Appendix Table 3.

As far as count status, pitches are marked if they are thrown in a three-ball or two-strike count. The interaction term in the model will account for the possibility of a relationship between the two. Appendix Table 4 showcases the amount of pitches thrown in either situation.

Inferential Results

To test the extent of the relationship between the type of pitch about to be thrown and factors such as the velocity and location of the previous two pitches, we ran a logistic regression model using 10-fold cross validation to simultaneously train and test our model. Using backwards elimination based upon AIC, we ran the models step-by-step. To demonstrate the steps taken, Appendix Table 5 showcases the forward selection process of Michael Wacha's model leaving the final model shown in Appendix Table 6.

Notably, the velocity of the second-most recent pitch is not significant by itself, but since the interaction with handedness is significant, the term is required to maintain a hierarchical model. The same is true with distance above the center for the pitch before. The overall model (as shown in Appendix Table 5) has an AIC of 425.29. From our 10-fold cross validation, we know that the model has a predictive accuracy of 67.533%. This number, however, is ever changing and depends on how the sample is grouped during the validation.

Considering the number of predictors in the full model, the steps taken to select the best regression equation for both Ryu and Cole are not shown here but were performed in the same manner as the equation for Wacha's. The final models for both are shown in appendix tables, with Cole's model having an accuracy of 58.977% and Ryu's model with an accuracy of 63.323% (see Appendix Tables 7 and 8).

The assumptions for the logistic regression model were satisfied as the response variable is binary, every observation is independent, and no issues were found with collinearity between the predictors.

Discussion

Wacha's model was the most accurate in its predictions, meaning by this criterion he is the most predictable pitcher out of the three. Despite having access to the same variables, the accuracies of the models are consistently different.

For each pitcher, a two-strike situation had a significant impact. This is the only significant variable in common throughout the three pitchers. Interestingly, it results in a different effect for Wacha than it does for Ryu or Cole, with Wacha's promoting more fastballs on two-strike counts.

Only Cole had the horizontal location of previous pitches show significance. However, Wacha and Ryu both significantly relied on previous vertical locations. All three models did have significant interactions using handedness, validating the claim that pitchers approach a left-handed batter differently than a right-handed batter. Because the batter can see the ball for slightly longer out of a pitcher's hand if they are of the opposite handedness, pitchers are often forced to pitch more carefully in these situations.

Velocity did not have a significant impact on Cole's model, while it did for both Ryu and Wacha. This makes sense considering Cole has a higher fastball velocity, meaning he can overpower hitters as opposed to relying on keeping them off-balanced with off-speed pitches.

If pitching predictability is a factor in success, these results make a lot of sense. Fielder Independent Pitching (FIP) is a measurement of success over a season. It is essentially an adjusted form of Earned Run Average (ERA) but considers other fielding factors to isolate the performance of the pitcher. The lower the FIP value, the better. Cole had an MLB leading FIP over the 2019 season of 2.64 and Ryu had an above average FIP of 3.10. Wacha, however, had a FIP of 5.61, which is significantly worse than the league average of 4.20.

Because this paper was only able to perform an analysis on three pitchers, we were unable to identify a large-scale relationship between FIP and predictability. Interestingly, as demonstrated in the results section, both Wacha and Ryu had a similar pitch arsenal and velocity on each pitch. Yet Ryu was able to perform much better in 2019 than Wacha, getting an invite to the All-Star Game and finishing second in the Cy Young race. While it is highly unlikely that low pitch predictability is the sole factor behind his success, it is hard to deny its influence. While the accuracies of the models are not extremely precise, the purpose of this study is to provide a simplistic way to observe pitch sequencing.

Despite the model being likely too complicated to use during a live game, it could still give many implications at the major league, minor league, and collegiate levels. Having information on what a pitcher is likely to throw in a two-strike or three-ball count could help a batter better prepare for games. Or he could use the information regarding sequencing, looking at both the location and velocity effects.

Additionally, sequencing is a commonly ignored factor when evaluating a pitcher. Front offices, when determining the value of a free agent, commonly look at statistics such as wins above replacement (WAR) as well as the trend of common statistics over the last several seasons. While these are undeniably important, there is no popular, quantitative way to measure the sequencing IQ of a pitcher. Given that the biggest limitation of this study was the number of pitchers used, for future research, it would be interesting to run these sets of individual models on a larger scale, mainly increasing the number of starting pitchers observed. This information would allow us to draw conclusions about the effect of pitch predictability, and possibly allow both major league teams and fans to evaluate pitchers with a higher degree of accuracy.

Works Cited

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference. *Sociological Methods & Research*, *33*(2), 261–304. doi: 10.1177/0049124104268644

Burris, K., Vittetoe, K., Ramger, B., Suresh, S., Tokdar, S. T., Reiter, J. P., & Appelbaum, L. G. (2018). Sensorimotor abilities predict on-field performance in professional baseball. *Scientific Reports*, *8*(1). doi: 10.1038/s41598-017-18565-7

Elfrink, T. (2018). Predicting the outcomes of MLB games with a machine learning approach *Business Analytics Research Paper*

Fernandes, C. J., Yakubov, R., Li, Y., Prasad, A. K., & Chan, T. C. (2019). Predicting plays in the National Football League. *Journal of Sports Analytics*, 1–9. doi: 10.3233/jsa-190348

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1 - 26. doi:http://dx.doi.org/10.18637/jss.v028.i05

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry L., and Müller, K. (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

Data. (n.d.). Retrieved from http://www.brooksbaseball.net/about.php

Appendix

Table 1:
Pitch Type Classification Breakdown

Pitcher Last Name	Total Number of Pitches	Classification Considered Fastball	Number Considered Fastball	Classification Considered Off-speed	Number Considered Off-speed
Wacha	726	Four-seam fastball Two-seam fastball Cutter	475 (65.4%)	Changeup Curveball Slider	251 (34.6%)
Ryu	854	Four-seam fastball Two-seam fastball Cutter	521 (61.0%)	Changeup Curveball Slider	333 (39.0%)
Cole	841	Four-seam fastball Two-seam fastball Cutter Sinker	457 (54.3%)	Knuckle-curve Slider Changeup	384 (45.7%)

Table 2
Logistic Regression Equation Variables

Variable Name in Equation	Variable Name	Description
Y_i	Pitch type	A binary categorical variable that describes pitch type. It can take on two values: fastball (0) and off-speed (1).
Z_1	Stance	A categorical variable representing the handedness of the batter. It is 0 if the batter is left-handed.
Z_2	Ball 3	A binary categorical variable that is equal to one in a situation where the hitter is in a three-ball count.
Z_3	Strike 2	A binary categorical variable that is equal to one in a situation where the hitter is in a two-strike count.
X_1	Velocity of the pitch before	A quantitative variable representing the velocity of the most recent pitch.
X_2	Velocity of the pitch two before	A quantitative variable representing the velocity of the second-most recent pitch.
X_3	Distance above before	A quantitative variable representing the distance of the pitch before in feet above the vertical center of the strike zone.
X_4	Distance below before	A quantitative variable representing the distance of the pitch before in feet below the vertical center of the strike zone.
X_5	Distance above two before	A quantitative variable representing the distance of the second-most recent pitch in feet above the vertical center of the strike zone.
X_6	Distance below two before	A quantitative variable representing the distance of the second-most recent pitch in feet above the vertical center of the strike zone.
X_7	Distance outside before	A quantitative variable representing the distance of the pitch before away from the direction of the batter from the middle of home plate.
X_8	Distance inside before	A quantitative variable representing the distance of the pitch before in the direction of the batter from the middle of home plate.
X_9	Distance outside two before	A quantitative variable representing the distance of the pitch two before away from the direction of the batter from the middle of home plate.
X_{10}	Distance inside two before	A quantitative variable representing the distance of the pitch two before in the direction of the batter from the middle of home plate.

Table 3
Pitch Location Summary

Pitcher Last Name	**Horizontal Location in feet** Measured with right side of plate being positive	**Standard deviation**	**Vertical Location in feet** Measured from the ground up	**Standard deviation**
Wacha	0.034	0.77	2.31	1.06
Ryu	0.132	0.83	2.42	0.89
Cole	0.010	0.78	2.40	0.99
Total	0.060	0.79	2.37	0.98

Table 4
Count Status Summary

Pitcher Last Name	**Pitches Thrown in 2 Strike Counts**	**Pitches Thrown in 3 Ball Counts**
Wacha	233 (32.09%)	70 (9.64%)
Ryu	242 (28.34%)	52 (6.09%)
Cole	287 (34.13%)	62 (7.37%)

Table 5
Wacha's forward selection process

Predictor Removed	**AIC**
No predictors removed	445.39
Pitch outside 2 before*handedness	443.39
Velocity before* handedness	441.42
Velocity before	439.49
Velocity before* handedness	437.57
Distance below 2 before* handedness	435.70
Pitch inside 2 before distance inside* handedness	433.92
Pitch inside 2 before	432.10
Pitch inside before* handedness	430.41
Ball 3* handedness	429.01
Ball 3	427.79
Pitch outside before	426.83
Pitch inside before	426.33
Distance below 2 before	426.22
Distance below before* handedness	425.92
Distance below before	425.29

Table 6
Wacha's final regression equation

Coefficient	Estimate	Standard Deviation	Z Value	P-Value
Intercept	-1.088	2.211	-0.492	0.623
Handedness	-9.030	3.794	-2.414	0.016
Strike 2	-0.659	0.336	-1.963	0.049
Velocity 2 before	0.008	0.025	0.343	0.732
Distance above before	0.400	0.281	1.423	0.155
Distance above 2 before	0.660	0.333	2.482	0.013
Velocity 2 before* handedness	0.089	0.042	2.121	0.034
Distance above before* handedness	-0.633	0.400	-1.581	0.114
Distance above 2 before* handedness	-0.859	0.440	-1.950	0.051
Strike 2*handedness	1.557	0.483	3.226	0.001

Table 7
Ryu's final regression equation

Coefficient	Estimate	Standard Deviation	Z Value	P-Value
Intercept	-28.703	7.676	-3.739	0.000
Handedness	25.396	7.852	3.235	0.001
Ball 3	-0.793	0.365	-2.172	0.030
Strike 2	1.141	0.643	1.774	0.076
Velocity before	0.310	0.085	3.641	0.000
Distance below before	2.130	0.744	2.861	0.004
Velocity*handedness	-0.273	0.087	-3.130	0.001
Distance below before*handedness	-1.780	0.769	-2.313	0.020
Strike 2*handedness	-1.473	0.860	-2.167	0.030

Table 8
Cole's final regression equation

Coefficient	Estimate	Standard Deviation	Z Value	P-Value
Intercept	-0.554	0.217	-2.551	0.011
Handedness	0.053	0.221	0.242	0.809
Ball 3	-1.167	0.320	-3.657	0.000
Strike 2	0.610	0.208	2.936	0.003
Pitch inside before	0.563	0.270	2.084	0.037
Pitch inside 2 before	-0.482	0.383	-1.257	0.209
Pitch inside 2 before* handedness	1.056	0.581	1.818	0.069