Mathematics and Sports

A Linear Regression Model for Predicting Whiff Percentage in Major League Baseball

Chris Greve[1] and Ryan Savitz[2,*]

[1] School of Business, Neumann University, 1 Neumann Drive, Aston PA 19014, [2] 2School of Business, Neumann University, 1 Neumann Drive, Aston PA 19014.

*Corresponding Author E-mail:
savitzr@neumann.edu

Abstract
This paper examines data related to the whiff percentage of five different Major League Baseball (MLB) pitches. A linear regression model is used to predict whiff percentage. Our results show that the models for each pitch, except the curveball, have statistical significance. The results of the cutter model are especially significant and give an indication of which pitchers in the MLB should throw their cutter more. The results found herein not only add a piece to the story but, also, lead to future areas of research in pitch modeling.

Keywords: whiff percentage, linear regression, cutter, MLB baseball, pitchers

1 Introduction

In Major League Baseball (MLB), pitchers utilize different grips and wrist rotations to create movement on the baseball as they throw it. There are different names for each type of pitch. The most commonly used pitches in the MLB are fastballs, curveballs, changeups, sliders, cutters, and sinkers. All these pitches move in different ways at different speeds. The way the baseball moves and its speed could be a factor in predicting future success. In this paper, data on sinkers, changeups, curveballs, sliders, and cutters were collected, and the expected whiff percentage was predicted.

Almost all pitchers in the MLB use a fastball which is the pitch that has the highest velocity and least amount of movement. All other pitches thrown are designed to move using speeds and movements, compared to the standard fastball. One of these other pitches is the sinker, which is a form of a fastball that with an altered grip has more downward vertical movement and/or strong-arm side movement than the average fastball [1,2]. Curveballs and changeups, on the other hand, move at a slower velocity and drop vertically as they cross the strike zone. Sliders move faster than curveballs but significantly slower than fastballs and move with high amounts of horizontal movement. Sliders can be differentiated from sweepers (which are not included in this analysis) by their lack of extreme horizontal movement[1]. Last, the cutter is a form of a fastball that moves away from the pitcher's arm side as it crosses home plate. The aim of this paper is to develop a model that will predict a pitcher's whiff percentage for the cutter pitch, and then to compare and contrast this model within the context of the other pitch types. The reason why we focus on the cutter is because it stands out as a unique pitch in baseball. It balances between the velocity of a traditional fastball and the slower pace of an off-speed pitch. Its unique blend of speed and movement intrigued us, and we wanted to quantify what makes a great cutter different from an average one.

Cam Rogers from FanGraphs did research related to this recently [6]. In Rogers' research, pitches were grouped into categories, so he did not ascertain an expected whiff percentage for individual pitch types. Rogers also included hitter variables such as the hitter's whiff rates, the strike count when the pitch was thrown, and more factors that were not included in our model. The variables that Rogers included that we did not include were not specific to individual pitch models and instead were situationally based. We, on the other hand, propose a model that is not in any way related to the batter that the pitcher is facing. Rogers for his final results combined all his pitch groupings together into one expected whiff percentage for each pitcher and it had a statistically significant R^2 value of 0.83.

Additionally, Eric Martin did research using similar variables to predict strikeouts instead of whiff percentage [5]. He used the difference in pitchers' movement and speed from their other pitches to forecast strikeout rates. Martin's model was statistically significant at predicting strikeouts with the R^2 value of 0.23. We now extend the research of Rogers (2020) and Martin (2019) by developing models to predict the whiff percentage for various types of pitches. As will be seen, the R^2 of our model is only slightly higher than Martin's, and well short of Rogers'. Rogers' model may have better predictive power due to his use of situationally based variables. We also, at the end of this paper, note some areas where our current model may be refined for future research.

2 Methodology

All of the baseball data used to create the models were found on baseballsavant.com [3]. Data can be accessed under the pitching arsenal stats and pitch movement tab. Our sample consists of data from 124 MLB pitchers. These pitchers were chosen for the sample since they all threw at least 50 cutters in the 2022 season.

The dependent variable in this model is the whiff percentage. Whiff percentage is calculated by taking the number of swings and misses divided by swings:

$$\textit{Whiff \% = (number of swings and misses)/(total number of swings)} \quad \textbf{(Eq. 1)}$$

We began the process of developing our models by focusing on the cutter pitch. We first considered the independent variables of spin rate, velocity of the pitch, horizontal movement, vertical movement, and total movement. Following a period of exploratory data analysis, we found that, somewhat surprisingly, the variables of horizontal movement and vertical movement had little predictive power, but that the sum of these variables did appear to be an important predictor of whiff percentage.

We also found at this time that pitch velocity did not appear to have the predictive power we expected, but that the difference between the velocity of a pitcher's fastest pitch and the velocity of the pitch under consideration did appear to be closely related to whiff percentage. Note that when we discuss the "difference in velocity," we are referring to the difference between the average speed of the pitcher's fastball and the average speed of their other pitch.

Hence, we constructed a linear regression model to predict the whiff percentage on cutter pitches using the independent variables of pitch velocity, the difference between the velocity of the pitcher's fastest pitch and the pitch under consideration, and total pitch movement [4].

Following this analysis, we used the three aforementioned independent variables to predict the whiff percentage for four additional types of pitches: sinker, changeup, curveball, and slider.

Following the construction of these regression models, we also verify the assumptions of regression by checking our models for the presence of multicollinearity, heteroskedasticity, and non-normality of residuals. We utilize the 0.05 level of significance throughout this research.

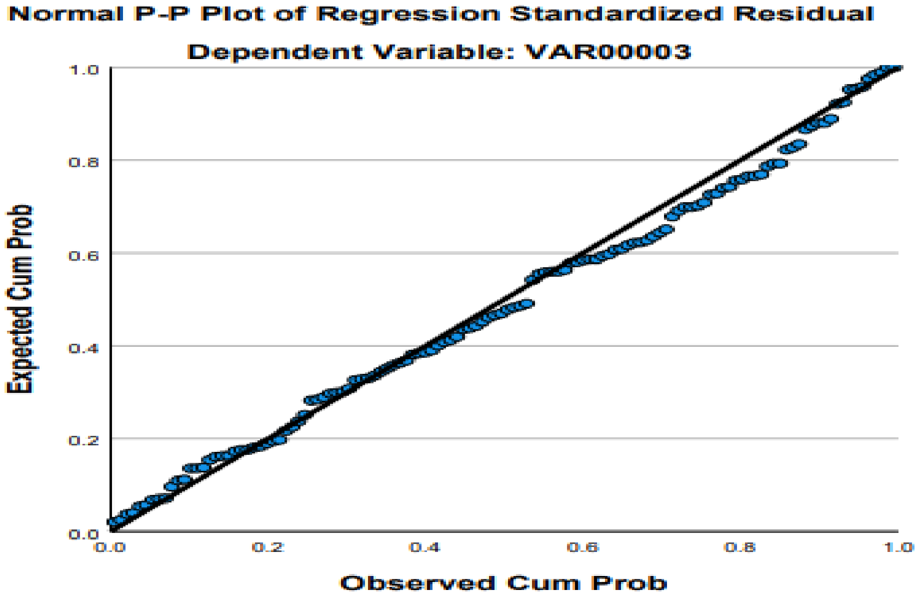

Figure 1: Normal Probability Plot for Cutter Pitch

3 Results

Following the proposed methodology, 5 separate linear regressions were performed for the following pitch types: sinker, changeup, curveball, slider, and cutter. The dependent variable (DV) in each model is whiff percentage for the given pitch type. The independent variables (IV) that were retained in the various models include the following: (PSpeed) pitch speed (MPH), (MPHDiff) the difference between the pitcher's fastest pitch and the pitch under consideration (MPH Difference), and (TotalMove) the total movement of the pitch in inches (vertical movement plus horizontal movement). This variable is called Total Movement (Inches). The results of these regressions are presented in Table 1, below. Note that, in this table, "P-value #" is the p-value for IV number #. Similarly, "VIF #" is the variance inflation factor for IV number #.

Table 1	Sinker	Changeup	Curveball	Slider	Cutter
DV	Whiff %	Whiff %	Whiff %	Whiff %	Whiff %
PSpeed	MPH	MPH	MPH	MPH	MPH
MPHDiff	MPH Difference	MPH Difference	MPH Difference	MPH Difference	MPH Difference
TotalMove	Total Movement (Inches)	Total Movement (Inches)	Total Movement (Inches)	Total Movement (Inches)	Total Movement (Inches)
P-Value 1	.001	.112	.268	.001	.001
P-Value 2	.663	.003	.526	.001	.011
P-Value 3	.828	.783	.472	.803	.001
VIF 1	1.322	1.976	2.954	2.77	2.067
VIF 2	1.146	1.742	2.926	2.315	2.861
VIF 3	1.249	1.23	2.98	2.705	2.014
R^2 Value	.118	.062	.061	.153	.273
P-Value for model's overall F-test	.002	.026	.194	<.001	<.001

Table 1: Regression Model Results

The results in table 1 indicate that only one pitch has statistical significance for all three independent variables. That pitch is the cutter. This is not surprising, as the cutter was the original pitch we attempted to model and, and the other four models are based on the original cutter model.

The cutter model had an R^2 value of 0.273 which means that this model explains 27.3% of the variability in whiff percentage. This compares favorably to Martin's (2019) R^2 for predicting strikeouts. The p-value for the overall F-test for this model was < 0.001, which indicates a high degree of statistical significance.

No concerns regarding the assumptions of regression were found to exist for this model. Specifically: (1) the normal probability plot was very close to a perfectly straight line with slope of unity; (2) a residual plot indicated no presence of heteroskedasticity; and (3) all of the variance inflation factors were well below 5, thus giving no indication of serious multicollinearity. Figure 1, below, shows the aforementioned normal probability plot, while figure 2 shows the residual plot.

While we did not expect to see the presence of heteroskedasticity or non-normality of residuals, it was reassuring to ascertain that multicollinearity was not a problem. Our original concern was that the use of the independent variables of pitch velocity and differential in pitch velocity could introduce a multicollinearity issue, but that did not turn out to be the case. We pause here to note that the models for the other pitches were also devoid of violation of regression assumptions.

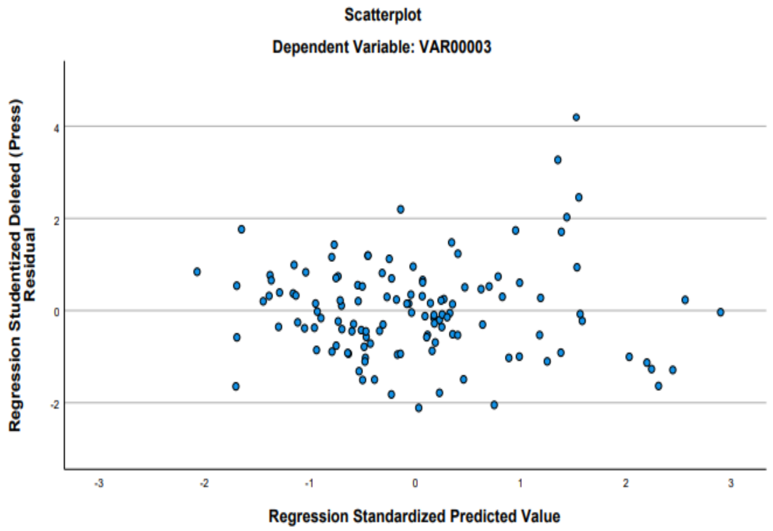

Figure 2 Residual Plot for Cutter Pitch

The regression equation for expected whiff percentage is as follows:

$$\textit{Expected cutter whiff percentage} = 1.055X1+.448X2+1.437X3 -89.661 \quad \textbf{(Eq. 2)}$$

where X1= Cutter velocity, X2= difference between velocity of fastest pitch and cutter mph, and X3= cutter total movement in inches.

Coupled with the results presented in table 1, we can see that a statistically significant, positive relationship exists between each of three independent variables and whiff percentage. Furthermore, the model as a whole has statistically significant predictive ability, as evidenced by the result of the F-test in table 1, above.

We now examine the regression equations for the non-cutter pitches. As with the cutter model, none of these models exhibit any issues regarding multicollinearity, non-normality of residuals, or heteroskedasticity. Note that the only model where all independent variables were statistically significant was the cutter model. We present the remaining models, as they are, so they can be easily compared to the cutter model. These comparisons will be discussed, in detail, in the following section. In the remaining models, the statistically significant independent variables are presented in bold.

$$\textit{Expected slider whiff percentage} = 1.341X1+1.625X2-0.024 -94.083 \quad \textbf{(Eq. 3}\textit{)}$$

where X1= slider velocity, X2= difference between velocity of fastest pitch and slider mph, and X3= slider total movement in inches. In this model, only pitch velocity and velocity differential are significant, although the model as a whole is statistically significant.

$$\textit{Expected curve whiff percentage} = 0.570+0.371X2-0.135X3 -10.550 \quad \textbf{(Eq. 4)}$$

where X1= Curve velocity, X2= difference between velocity of fastest pitch and curve mph, and X3= curve total movement in inches.

This model contains no statistically significant variables, and the model as a whole lacks any significant predictive ability

$$\textit{Expected change up whiff percentage} = 0.487X1+1.246X2-0.040X3 -19.148 \quad \textbf{(Eq. 5)}$$

where X1= change up velocity, X2= difference between velocity of fastest pitch and change up mph, and X3= change up total movement in inches.

In this model, only the velocity differential is statistically significant, although the model as a whole has significant predictive ability.

$$\textit{Expected sinker whiff percentage} = 0.656X1+0.145X2+0.015X3 -46.39 \quad \textbf{(Eq. 6)},$$

where X1= sinker velocity, X2= difference between velocity of fastest pitch and sinker mph, and X3= sinker total movement in inches.

Pitch velocity is the only significant variable in this model, and the model as a whole does have significant predictive power.

4 Discussion and Conclusion

We will first examine the real-life value of the model we constructed for the cutter pitch. The best way to loom at this is by using real pitcher data. For example, consider Cal Quantrill of the Cleveland Guardians, who threw the second most cutters of all MLB pitchers in 2022. To start, Quantrill's cutter on average was 88.5 miles per hour so that will be the value of X1 in the equation. Next, Quantrill's fastest pitch was his fastball at an average of 93.4 miles per hour. His fastball speed of 93.4 miles per hour subtracted by his cutter speed of 88.5 miles per hour yields a 4.9-speed differential. As a result, 4.9 will be the value of X2 in the equation. Finally, Quantrill had 30.5 inches of total movement on his cutter and this will be the value of X3. Plugging these values into equation (2) yields an expected whiff percentage of 24.41%. This can be compared to Cal Quantrill's actual whiff percentage of 23.8, which appears to be a relatively close estimate. We pause to note here that the mean standard error of prediction for our model was found to be 1.140.

This type of result is not the case for all players, however. Some pitchers may have a much higher expected whiff percentage than regular whiff percentage and vice versa. This could due to error in the model, or something special about these pitchers. Examining this more closely, we refer to table 2 below, which shows the top 5 pitchers in MLB in cutter expected whiff% and how much they throw their cutter.

Pitcher	Cutter Expected Whiff Percentage	Cutter Whiff Percentage	Cutter usage
Framber Valdez	35.48%	35.3%	10.27%
Sonny Gray	34.11%	35.6%	9.14%
Spenser Watkins	33.64%	25.5%	27.89%
Bryan Baker	33.09%	22.5%	26.42%
Spencer Howard	32.82%	24.6%	35.26%

Table 2: Expected vs Actual Whiff Percentage

From table 2 it appears Framber Valdez and Sonny Gray have great cutters that they may not be fully utilizing. Framber Valdez has the best-expected whiff percentage on cutters in the entirety of MLB, and his expected whiff percentage matches his actual whiff percentage very closely. Of all of Valdez's pitches, his cutter's expected whiff percentage is better (higher) than the actual whiff percentage of all his other pitches. Valdez only throws the cutter 10.27% percent of the time. With more usage of this pitch, Framber could potentially see an increase in strikeouts next season.

Next, Sonny Gray has never used his cutter much over his entire career. In 2020 he didn't throw his cutter at all. Then he brought it back in 2021 and added 19.1 inches of movement. Gray is still new to throwing his revamped cutter and may not feel fully comfortable throwing this pitch yet. However, it may be Gray's best pitch. His cutter has a slightly better expected whiff than all of his other commonly used (n >= 100). Gray may not know how dominant his cutter is yet and more usage of this pitch can help him break out. The other three pitchers in the top 5 all throw their cutter about a quarter of the time. Their actual cutter whiff percentages, however, are noticeably lower than their predicted cutter whiff percentage. The reason that the whiff percentages of Watkins, Baker, and Howard are higher than their actual whiff percentage could be because they throw the pitch more frequently than other pitchers so batters may see it coming. The pitchers' teams have to attempt to find the perfect balance where they are throwing their cutter an effective amount but, simultaneously, do not overuse it, thus making it easily recognizable by batters. For these top pitchers, it may be beneficial to use this pitch in high-leverage situations, such as when the batter has two strikes, to try to generate more strikeouts.

All of the other pitches in this paper, besides the cutter, have at least one independent variable that is not statistically significant. It is possible, however, that using a larger sample size in future research (which includes several seasons of data) could result in the presence of mores statistically significant variables. In sinkers, the miles per hour difference and total movements are statistically insignificant. The reason the miles per hour difference did not have significance for sinkers could be attributed to the fact sinkers on average were thrown only 0.28 MPH slower than fastballs, thus making this velocity differential trivial. The cutter, however, was on average 5.06 MPH slower than the fastball.

Regarding changeups, the velocity, and total movement variables are statistically insignificant. Changeups are a high-risk, high-reward pitch and their success is dependent on the location it is thrown and the situation. Unlike the cutter, the speed and movement of the changeup do not have much of an impact if the pitch is thrown right down the middle of the plate. For sliders the total movement variable was insignificant. This may also be due to the risk-reward nature of this pitch. If a pitcher has a fast-moving slider, they can get away with throwing it across the middle of the plate since it isn't always a full off-speed pitch. However, if a pitcher throws their slider as slow as a changeup or curveball and right down the middle it will likely be hit regardless of how much movement it has.

The reason the curveball model failed may have been because of the significant difference between the curveball and the baseline pitch of the cutter. The curveball typically is a huge risk-reward pitch for pitchers. If a pitcher lets their curveball hang it will normally get hit regardless of its speed and movement. On the other hand, the cutter can be left over the middle of the plate and not be beaten as easily if it has the right amount of velocity and movement. The variables selected for this model best fit the cutter as it is a unique pitch that is in between a fastball and off-speed pitch in terms of speed so the factors that make it successful are different than any other pitch.

The model constructed for the cutter herein can be helpful for baseball organizations. It can be used to give a fairly accurate estimate of what a pitcher's whiff percentage will be over time, as opposed to simply using guesswork and anecdotal evidence. Additionally, the results of the model can help pitchers to determine how often their cutter is worth throwing in games. Finally, the model can be used as a tool for pitchers to use in the off-season. Using the example of Cal Quantrill, who has an 88.5 MPH cutter, with a 4.9 MPH speed differential, and total movement of 23 inches equaling an expected whiff percentage of 22.41 the model can predict how his whiff percentage might change if he added or lost speed and movement. If Quantril added one inch of movement so his total movement is 24 inches his expected whiff percentage would go up to 22.86. This is useful data for pitchers to use when figuring out how they need to improve. Hence, during the off-season, pitchers might experiment with different combinations of velocity and movement, and see which combination of these variables (that they are physically capable of) yield the highest expected whiff percentage.

Overall, the cutter model has statistical significance but it also has limitations. First, there were a few outliers of pitchers who had significantly higher whiff percentages than the expected whiff percentages. Table 3 shows the three largest outliers.

Pitcher	Cutter Whif Percentage	Cutter Expected Whiff Percentage
Jose Alvarado	55.7%	29.91%
Adrian Sampson	50%	29.20%
Huascar Brazoban	45.5%	30.01%

Table 3: Whiff Percentage Outliers

The cause for why these pitchers expected whiff percentage and actual whiff percentages are so far apart could be due to how the pitchers use their cutter. All three pitchers have at least a 5.8 mile per hour difference or more on their cutter to their fastest pitch. The model does account for the speed difference but with extreme examples like these three pitchers, but, naturally, prediction is going to be most accurate when the values of the variables are close to their respective means. These pitchers also all use their cutters differently than other pitchers. Since they all have a significant MPH difference their cutter is more deceiving to batters. These pitchers' cutters are used more as an off-speed pitch which generally generates a higher whiff%.

Given the results of the model, there is still room for improvement to explain more of the variability in whiff percentage. Some options for additional independent variables that can be added are a quadratic term for speed differential, location of the pitch, situation, and the batter at the plate. That said, we must be careful to maintain a relatively parsimonious model, like our current one, as we want the model to be easy to use and apply.

An additional area for future research would be constructing whiff percentage models, from scratch, for the curve, sinker, changeup, and slider. While this paper examines the relationship between the cutter and the other pitches, purely individualized models may have better results regarding practical applications.

In baseball, various factors play a role in determining the outcome of a game. Utilizing statistics such as these is crucial in analyzing and forecasting the performance of teams and players. We hope this research provides avenues for future work, and that the models presented herein can be improved. One of the great things about baseball is that there are so many factors at play in any given game, resulting in a great deal of uncertainty. Models like these can allow us to better understand and utilize one specific aspect of this great game.

5 Limitations and Areas for Future Research

A few limitations, which give rise to areas for future work, must also be noted regarding this research. The authors kindly thank the anonymous reviewers for some of these ideas. First, there may be additional, related independent variables that can be used to help predict whiff percentage. One of this would be to replace the total vertical plus horizontal pitch movement with the Euclidean distance of the pitch movement. It is also interesting to note that this would affect some pitchers (those with relatively equal horizontal and vertical movements) more than others (those whose pitch movement is primarily confined to one dimension).

This leads us to another limitation to note: the analysis we performed did not control for either pitcher or hitter. It could be interesting to create a fixed effects model to control for pitcher and/or somehow control for hitter, and then to compare these results with those presented herein. This comparison could add additional insight to complex physical phenomenon of throwing and attempting to hit a pitch.

References

1. Castrovince, Anthony. A Fan's Guide to Baseball Analytics: Why WAR, WHIP, wOBA, and Other Advanced Sabermetrics Are Essential to Understanding Modern Baseball. Simon and Schuster, 2020.
2. “Pitch Types: Glossary.” MLB.com. Accessed December 3, 2024. https://www.mlb.com/glossary/pitch-types.
3. “Baseball Savant: Statcast, Trending MLB Players and Visualizations.” baseballsavant.com. Accessed January 26, 2023. https://baseballsavant.mlb.com/about
4. Adam, H., K. Khadija, and K. Suzanne. "Stepwise Regression: Definition, Uses, Example, and Limitations." Investopedia (2022).
5. Martin, Eric P. "Predicting Major League Baseball Strikeout Rates from Differences in Velocity and Movement Among Player Pitch Types." In MIT Sloan Sports Analytics Conference. 2019.
6. Rogers, Cam. “Introducing Probabilistic Pitch Scores and xWhiff Metrics.” Community Blog, October 7, 2020. https://community.fangraphs.com/introducing-probabilistic-pitch-scores-and-xwhiff-metrics/.