FORECASTING BATTER PERFORMANCE USING STATCAST DATA IN MAJOR LEAGUE BASEBALL

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Nicholas Christopher Taylor

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Applied Statistics

April 2017

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Forecasting Batter Performance Using Statcast Data in Major League Baseball

By

Nicholas Christopher Taylor

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Rhonda Magel
Chair

Edward Deckard

Seung Won Hyun

Approved:

04/10/2017
Date

Rhonda Magel
Department Chair

ABSTRACT

2015 saw the release of the Statcast camera system within Major League Baseball ballparks, which provided statisticians with new data to analyze. One statistic, average exit velocity, is of particular interest. We would like to see if a batter's average exit velocity can significantly explain the variation in his slugging percentage and batting average on balls in play (BABIP) when taken into account with other, more traditional baseball statistics. These two statistics are of particular interest within advanced baseball data analysis.

We found that a player's average exit velocity can significantly explain the variation in both his slugging percentage and his BABIP. We also discovered that the significance is stronger in explaining slugging percentage than in explaining BABIP.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BABIP.....batting average on balls in play

FB%.....fly ball percentage

FBProp.....fly ball proportion

LD%.....line drive percentage

LDProp.....line drive proportion

Hard%.....hard contact percentage

HardProp.....hard contact proportion

HR.....home run

Soft%.....soft contact percentage

SoftProp.....soft contact proportion

Spd.....speed score

LIST OF APPENDIX TABLES

Table Page

CHAPTER ONE: INTRODUCTION

The history of baseball is one of continuously improving player evaluation. During baseball's formative years, the player pool was tiny, consisting only of white males from the United States. Hall of Famer Jackie Robinson broke the color barrier in 1947, and today, players are scouted internationally as teams continue to look for sources of talent that can be acquired for sub-market prices.

In addition to utilizing broader pools of talent, teams and fans alike have always been searching for ways to assess player performance through statistical analysis. Bill James, one of the earliest pioneers of advanced baseball statistical analysis, coined the term 'sabermetrics' (after SABR, the Society for American Baseball Research) in 1980, describing it as "the search for objective knowledge about baseball" (Birnbaum). Traditional statistics, such as runs batted in (RBIs), batting average (BA), and saves have fallen by the wayside as methods of analyzing player performance. Nowadays, baseball fans and front offices alike are utilizing more objective, less noisy ways of predicting a player's true talent level – his skill as determined by him and not by a series of unpredictable circumstances. Statistics such as field independent pitching (FIP), batting average on balls in play (BABIP), weighted on-base average (wOBA) and wins above replacement (WAR) have entered the public sphere and are commonly cited as ways to assess a player's true performance level.

While statisticians have long had myriad game performance statistics available to them, sources of *physical* performance (such as the rotation rate of a pitcher's curveball, or a baserunner's time sprinting to first base) have not historically been available. However, 2015 saw the debut of Statcast. According to Major League Baseball, "Statcast, a state-of-the-art tracking technology, is capable of gathering and displaying previously immeasurable aspects of the game. Statcast collects the data using a series of high-resolution optical cameras along with radar equipment that has been installed in all 30 Major League ballparks. The technology precisely tracks the location and movements of the ball and every player on the field at any given

time." Statcast is able to track data such as pitch velocity, perceived pitch velocity, a pitch's spin rate, a batted ball's exit angle and exit velocity off the bat, a fielder's route efficiency to a ball, and myriad other physical feats (Casella). While the full data set has not yet been made available to the public (Berg), batted ball exit velocity, generally one of the most popularly cited Statcast statistics, is available in relative abundance through the website *Baseball Savant*. (Willman). For this research, we will examine the relationship between a player's slugging percentage and his average exit velocity. We will also examine the relationship between a player's batting average on balls in play (BABIP) and his average exit velocity. In addition, models will be developed to help explain a player's BABIP and slugging percentage based on his average exit velocity and five other variables. These variables are hard contact proportion (hardProp), soft contact proportion (softProp), line drive proportion (LDProp), fly ball proportion (FBProp), and speed score. These variables will be defined in chapter two. The rationale for including these six variables will also be discussed.

CHAPTER TWO: SURVEY OF LITERATURE AND DEFINITION OF TERMS

Before proceeding with an overview of the current state of research on average exit velocity, it's important to define the baseball terminology that will be used frequently in this thesis. Detailed explanations are given below; a table containing more concise definitions can be found in (Table A1). All definitions and formulas given below are taken from the Fangraphs Library ("Complete List…"), with the exception of average exit velocity, whose definition is taken from the Major League Baseball website ("Exit Velocity…").

Overview of Baseball Statistics used in this Thesis

Slugging percentage is measure of a batter's power, which is defined as his ability to hit for extra bases. Slugging percentage is an important skill because players who hit for power produce more runs for their team, and scoring more runs leads to more wins. For this reason, slugging percentage has long been held up as a prized and vital skill by sabermaticians. The formula for slugging percentage is given as:

$$\#Total\ Bases\ /\ \#At\ Bats$$

Where total bases is defined as:

$$Total\ Bases = \#1B + 2*\#2B + 3*\#3B + 4*\#HRs.$$

Where #1B=number of singles, #2B=number of doubles, #3B=number of triples, and #HRs=number of home runs. In this formula, one base is awarded for each single a batter hits, two bases for each double, three bases for each triple, and four bases for each home run. Because slugging percentage is found by dividing a player's total bases by his total number of at-bats, slugging percentage can be thought of as the number of total bases a player obtains per AB ("Complete List…"). A good slugging percentage is above 0.450 (Simon).

It's important to note that while the term slugging percentage contains the word 'percentage,' this is merely an etymological quirk. It is not actually a percentage, as it is theoretically possible to have a slugging percentage exceeding one, if a player has more total bases than at-bats. In fact, the theoretical upper bound of slugging percentage is actually four, if a player were to hit home runs in every single AB. Also, in this thesis, when we refer to a one point increase in slugging percentage, we are referring to an increase of 0.001.

Batting average on balls in play (BABIP) gives the rate at which balls that are put into play become hits. Any plate appearance that does not result in a walk, strikeout, hit by pitch, catcher's interference, sacrifice bunt, or home run involves a ball being put in play. The formula for BABIP is given as:

$$BABIP = (H - HR) / (AB - K - HR + SF)$$

Where H=hits, HR=home runs, AB=at bats, K=strikeouts, and SF=sacrifice flies. Roughly 30 percent of balls in play fall in for hits. A typical batter will have a BABIP between 0.270 and 0.330. Sabermaticians generally use BABIP as a forecasting tool, to determine whether or not a player performing above or below his career averages is getting lucky or unlucky. For example, a poor player who suddenly starts hitting for a high average might be getting lucky with the number of balls that fall in for hits, which would be reflected through a high BABIP. High BABIPs are generally unsustainable, and this player's performance is likely to drop once their BABIP normalizes. However, if we can identify player skills that have a predictive impact on BABIP, then we can demonstrate that BABIP is not purely luck-based for all players ("Complete List…").

It's important to note that BABIP, unlike slugging percentage, is a true proportion and cannot exceed one. Also, in this thesis, when we refer to a one point increase in BABIP, we are referring to an increase of 0.001.

Speed score (Spd) is a statistic developed by Bill James, one of the fathers of sabermetrics, which is composed of a player's stolen base percentage, frequency of stolen base

attempts, percentage of triples, and runs scored percentage. Speed score is measured on a scale from zero to ten, with ten being the fastest and zero being the slowest ("Complete List…"). We're interested in using speed score to predict BABIP, since a fast player is more likely to beat out slow infield ground balls than a slower player, which would make speed a true talent predictor of BABIP. An example of speed score increasing by one would be an increase from four to five.

Line Drive Percentage (LD%) is the percentage of a batter's balls in play that are line drives. Line drives are the most ideal batted ball result a player can produce, as they very often fall in for hits. In 2014, lines drives fell in for hits 68.5 percent of the time ("Complete List…"). As such, we believe that players with high line drive rates will have a higher BABIP than players with lower line drive rates. We also believe that players with a high percentage of line drives will have a higher slugging percentage.

Fly Ball Percentage (FB%) is the percentage of a player's balls in plays that are fly balls ("Complete List…"). Well-struck fly balls are the most likely batted ball type to become home runs, so we're interested in examining the relationship between a player's FB% and his slugging percentage.

The variables soft, medium, and hard contact percentage (soft%, medium%, and hard%), or the quality of contact percentages, very roughly, refer to the percentage of a player's batted balls that were struck with soft, medium, and hard contact. Whether or not a batted ball counts as soft, medium, or hard is determined by an algorithm developed by Baseball Info Solutions. The algorithm takes several factors into account, including the batted ball type (ground ball, fly ball, or line drive), landing spot, and hang time ("Complete List…"). It does not include exit velocity, which is why our goal is to also include average exit velocity to make predictions on a batter's performance. Because this algorithm is proprietary, we cannot examine the exact model used to categorize quality of contact, but we can still implement the variables into our analysis. We expect batters who make hard contact a high percentage of the time to perform better than batters who make a lot of soft and medium contact.

We are only including soft% and hard% in our analysis because, when written as proportions, soft%, medium%, and hard% will all add up to one. Because one of these variables can be written as a linear combination of the other two, including all three in the model would result in excessive variance inflation factors (VIFs) due to multicollinearity ("Multicollinearity…").

It's crucial to note that the variables LD%, FB%, soft%, medium%, and hard% are typically expressed on baseball statistics websites as percentages, but in our analysis, we use the proportion equivalents. Throughout the rest of this thesis, for the sake of simplicity, we will refer to these variables as LDProp, FBProp, softProp, and hardProp as a reminder to the reader that our analysis is working in terms of proportions, not percentages. Thus, an increase of one percent in any of these variables will correspond to an increase of 0.01, proportionally speaking.

Finally, a player's average exit velocity tells us, on average, how fast the ball travels in miles per hour when a batter puts it in play. This includes all batted ball events. Because a high average exit velocity indicates a player's ability to consistently make solid contact, we predict that players with a higher average exit velocity will have higher BABIPs and slugging percentages than players with lower average exit velocities ("Exit Velocity…").

Past Research

Ever since Statcast went live in 2015, sabermaticians have been making use of the new data made available. Most of this research takes the form of articles posted online by sabermaticians. Common sense tells us that batters that consistently strike the ball for a high exit velocity perform at a higher levels than batters with a lower average exit velocity. The research performed up to this point supports this hypothesis. For instance, May of 2015, baseball writer and researcher Rob Arthur published an article which discusses the impact of a player's average exit velocity on his OPS (on-base plus slugging percentage, which is a simple, commonly-used statistic that gives a snapshot of a player's offensive performance). For players

with more than twenty batted ball events at that point in the 2015 season, Arthur performed simple linear regression to predict OPS as a function of average exit velocity. He obtained a statistically significant R^2 value of .1475. While this may not seem like much, in a sport with as much noise as baseball, being able to explain 15Prop of the variance in OPS with a single statistic is not insignificant (Arthur). A fitted line plot of the regression model is given in (Fig. 1).

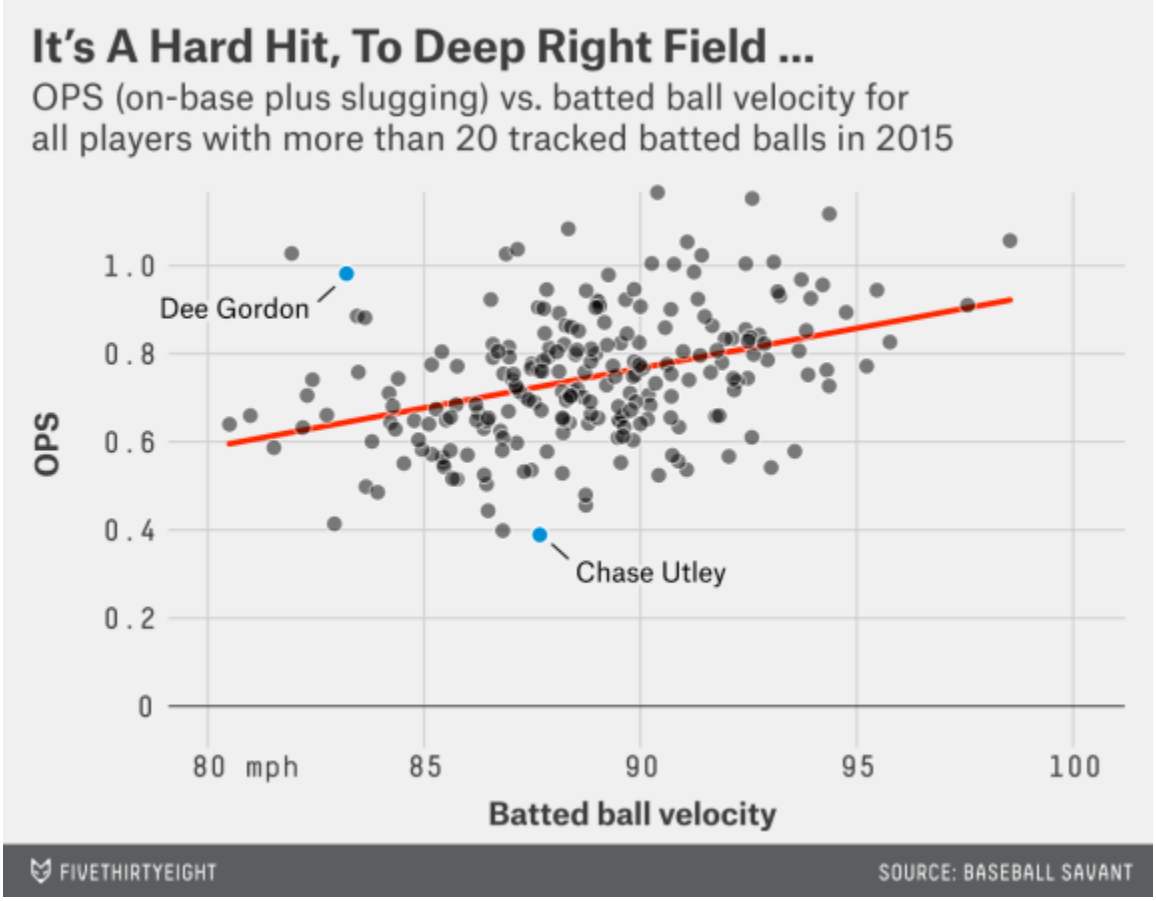

Figure 1. Fitted line plot predicting a batter's OPS from his average exit velocity using early 2015 season data. Arthur, Rob. "Chase Utley is the Unluckiest Man in Baseball." *FiveThirtyEight*. ESPN, 15 May 2015. Web

In a similar article published in July of 2015, sabermatician Stephen Shaw performed simple linear regression to predict a player's slugging percentage from his average exit velocity. He obtained an R^2 value of 0.29, meaning we can explain nearly 30 percent of the variation in slugging percentage through exit velocity alone (see Fig. 2) (Shaw). These results are encouraging, and it makes more sense to perform regression to predict slugging percentage

rather than OPS, as we don't necessarily expect a player's on-base percentage to be significantly influenced by his average exit velocity.

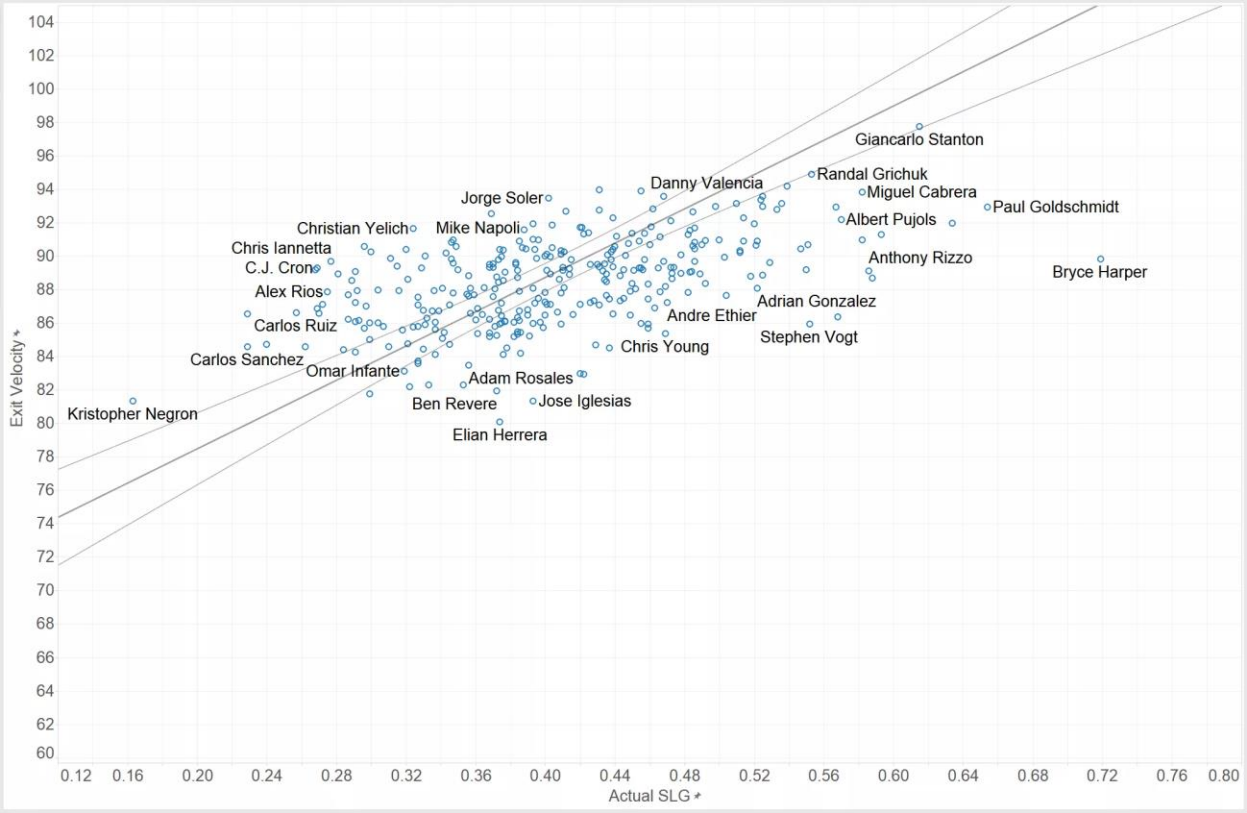

Figure 2. Fitted line plot predicting a player's slugging percentage from his average exit velocity using early 2015 season data. Shaw, Stephen. "Updated MLB Statcast Data (July 2015)." *Banished To The Bullpen*. Wordpress, 1 July 2015. Web.

Additionally, in September 2016, baseball researcher Billy Stampfl released an article that discusses a batter's expected performance based on his Statcast batted ball profile, which includes average exit velocity and launch angle off the bat. In investigating the correlation between average exit velocity and several offensive statistics, he discovered that the coefficient of determination when predicting slugging percentage is 0.3953 (see Table 1). In other words, nearly 40 percent of the variation in slugging percentage can be explained by average exit velocity alone, which is very large for a single predictor variable (Stampfl).

Table 1: Coefficient of determination between various offensive statistics.[a]

Variable 1 (Statcast)	Variable 2 (Fangraphs)	Correlation (R-squared)
Barrels/PA	wRC	0.4034
Barrels/PA	SLG	0.5900
Barrels/PA	BA	0.0021
Barrels/PA	wOBA	0.3970
Barrels/PA	HR/g	0.7513
Barrels/PA	ISO	0.76470
Avg. Exit Velocity	wRC	0.3173
Avg. Exit Velocity	wOBA	0.3336
Avg. Exit Velocity	SLG	0.3953
Avg. Distance	wRC	0.2440
Avg. Distance	wOBA	0.2698

Source: Stampel, Billy. "Barrels, Normative Analysis, and the Beauties of Statcast." *The Hardball Times*. The Hardball Times, 29 Sept. 2016. Web.

a. In the regression procedure, variable 1 is the dependent variable, and variable 2 is the independent variable.

Average exit velocity has value beyond predicting player performance. MLB published an article in November 2015 discussing how a player's average exit velocity can be an indicator of his health. Just as a pitcher losing velocity can be an indicator of a nagging injury, a batter posting a lower than usual exit velocity may also need a trip to the disabled list. For instance, from the beginning of the 2015 season through June of that year, Nationals first baseman Ryan Zimmerman posted an average exit velocity below 90 miles per hour. During this time, it was known that Zimmerman was attempting to play through a foot injury. His slugging percentage during this period was a paltry .265. On June 11th, Zimmerman was placed on the disabled list. From when he returned in late July through the end of the season, his average exit velocity

leaped to over 95 MPH, and his slugging percentage surged over a hundred points to .372 (Petriello).

Figure 3. Ryan Zimmerman's monthly average exit velocity before and after going on the disabled list. Petriello, Mike. "3 cool lessons from Statcast's debut season." *MLB.com*. Major League Baseball, 12 Nov. 2015. Web.

The research performed so far indicates that average exit velocity is positively correlated with batter performance, including slugging percentage. Thus, it would be useful for a team to examine a player's average exit velocity when considering whether or not to acquire him. It was also discovered that average exit velocity can be an indicator of player health, which would make it a valuable statistic to monitor if a player's performance begins to dip. This makes average exit velocity a useful tool when examining whether or not a scuffling player is fighting a lingering injury.

CHAPTER THREE: METHODS

The purpose of this research is to see how much variation in slugging percentage and BABIP can be explained by a player's average exit velocity when taken into consideration with several other intuitive statistics that we believe might also perform a significant role in explaining the variation in BABIP and slugging percentage. To achieve this purpose, ordinary least squares regression will be used to construct two models: one to explain a player's BABIP, based on a player's average exit velocity and other intuitive statistics, and the other to explain slugging percentage based on these same statistics. The independent variables, or intuitive statistics, initially considered in building these models are hardProp, softProp, FBProp, LDProp, speed score, and average exit velocity. Before we conduct the regression analysis, we will first generate correlation plots between each of the independent variables and the dependent variables, BABIP and slugging percentage, to examine the significance of the correlation between average exit velocity and the dependent variables, and how this correlation ranks among all independent variables. We will then run the regression procedure with all independent variables included, then drop any variables that are not significant at the alpha=0.10 significance level. Stepwise regression will be used to verify that we've selected the correct significant variables. The best subsets routine will be used, and models will be compared. We will consider dropping any significant variables that do not practically contribute to the explanation of variance in the dependent variable. Models will be compared on the basis of R^2, adjusted R^2, predicted R^2, Mallow's CP, the regression standard error, and variance inflation factors. We will be examining variance inflation factors because high levels of multicollinearity (which refers to high correlation between independent variables), can lead to inaccurate predictor coefficients. Because we would like to practically interpret the variable coefficients, model multicollinearity would be a problem for us ("Multicollinearity…"). The ordinary least squares regression assumption of normally distributed residuals with mean zero and constant

variance will also be checked. Based on these metrics, a final model will be recommended. We will follow the same procedure when developing a model to predict a player's BABIP.

The data used for developing the models will be taken from two online baseball statistics databases: *Baseball Savant* (Willman) and *Fangraphs (Fangraphs)*. From Baseball Savant, average exit velocity data will be gathered. From Fangraphs, data on hardProp, softProp, FBProp, LDProp, speed score, BABIP, and slugging percentage will be gathered. In order to avoid excess error due to small sample sizes, we will only be gathering data from players with at least one hundred batted ball events. All data taken will be from the 2015 Major League Baseball Season.

To validate the final models selected to explain variation in BABIP and slugging percentage, we will insert values of the significant independent variables from the 2016 Major League baseball season into the models. This will provide an estimation for a player's 2016 BABIP and slugging percentage. We will then compare the estimated BABIP obtained from the model for each player along with each player's actual BABIP, and R^2 and adjusted R^2 will be calculated. This will be compared with the original R^2 and adjusted R^2 values. The same thing will be done with the slugging percentage model. Data for the 2016 season was gathered from *Fangraphs* (*Fangraphs)* and *Baseball Savant* (Pullman). Data on hardProp, softProp, FBProp, LDProp, speed score, BABIP, and slugging percentage was collected from *Fangraphs*. Data on average exit velocity was collected from *Baseball Savant*. Again, we only collected data on players with at least one hundred batted ball events.

CHAPTER FOUR: RESULTS

Explaining variation in slugging percentage

The first model we built attempts to explain the variation in slugging percentage using our independent variables. We ended up selecting 2015 season data from 345 players with at least one hundred batted ball events. This data was used in both models. Because our dependent variable is slugging percentage, we first produced a correlation plot to examine the strength of the relationship between slugging percentage and average exit velocity, and we also examined the strength of the relationships between slugging percentage and each of the remaining independent variables in the model and compared these strengths to each other (Table 2).

Table 2: Correlation matrix between LDProp, FBProp, speed, slugging percentage, softProp, hardProp, and average exit velocity

	Slugging percentage	**Average exit velocity**	**LDProp**	**FBProp**	**Speed**	**SoftProp**
Avg. exit velocity	0.597					
LDProp	0.089	0.007				
FBProp	0.258	0.141	-0.297			
Speed	0.055	-0.086	-0.062	-0.153		
SoftProp	-0.365	-0.384	-0.355	-0.171	0.177	
HardProp	0.634	0.589	0.183	0.360	-0.164	-0.575

We see that the highest correlation with slugging percentage is hardProp, with r=0.634. This indicates a moderately strong positive relationship between slugging percentage and hardProp. The second most significant correlation with slugging percentage belongs with our variable of interest, average exit velocity, with r=0.597, also indicating a moderately strong

positive relationship. Squaring the correlation coefficient yields a coefficient of determination R^2 equal to 0.356. This means that, when considered on its own, average exit velocity can explain 35.6 percent of the variation in slugging percentage. We predicted that players with a higher average exit velocity will have a higher average slugging percentage, so these preliminary results are encouraging.

Our next step was to construct an ordinary least squares regression model to explain the variation in slugging percentage, taking each independent variables into account (see Tables 3 and 4).

Table 3: ANOVA output of full regression model predicting variation in slugging percentage.

Source	**DF**	**Adjusted SS**	**Adjusted MS**	**F-value**	**P-value**
Regression	6	0.76690	0.12782	58.92	0.000
LDProp	1	0.00379	0.00379	1.75	0.187
FBProp	1	0.01195	0.01195	5.51	0.019
Spd	1	0.04102	0.04102	18.91	0.000
SoftProp	1	0.00031	0.00031	0.14	0.707
HardProp	1	0.10966	0.10966	50.55	0.000
Avg - MPH	1	0.12145	0.12145	55.99	0.000
Error	338	0.73321	0.00217		
Total	344	1.50011			
Model Summary					
S	R^2	R^2 (adjusted)	R^2 (predicted)		
.04658	51.12%	50.26%	49.08%		

Table 4: Coefficient output from full model predicting variation in slugging percentage.

Term:	Coefficient	SE Coefficient	T-value	P-Value	VIF
Constant	-0.702	0.121	-5.82	0.000	
LDProp	0.1264	0.0956	1.32	0.187	1.43
FBProp	0.1028	0.0438	2.35	0.019	1.43
Spd	0.00678	0.00156	4.35	0.000	1.05
SoftProp	0.0325	0.0864	0.38	0.707	1.71
HardProp	0.5251	0.0739	7.11	0.000	2.29
Avg. exit velocity	0.00977	0.00131	7.48	0.000	1.63

Our full regression models produces a coefficient of determination value of 0.5112, which means that just over half the variation in slugging percentage can be explained by the independent variables in the model. Our adjusted and predicted coefficients of determination are 0.5026 and 0.4908 respectively, an insignificant reduction from our original R^2 value of 0.5112. Adjusted R^2 imposes a penalty for adding superfluous independent variables, and predicted R^2 indicates how well the regression model can predict new responses. Because our adjusted R^2 and predicted R^2 are close to our original R^2, this implies that the model does not have excessive independent variables, nor will it suffer a loss in accuracy when predicting new responses (Frost).

Examining the results of the coefficients table, we see that LDProp and softProp are insignificant at the alpha=.05 level. The p-value of softProp is 0.707, and the p-value of LDProp is 0.187. This implies that softProp is not significantly different from mediumProp, and can be removed from the model. HardProp is statistically significant ($p<.0005$), and meaningfully explains the variation in slugging percentage.

Surprisingly, we also notice that line drive percentage is not a significant predictor of slugging percentage (p-value = .187) when taken into account with the other independent variables. We speculate that while line drives can result in a single, double, triple, or home run, line drive rate alone is not enough to tell us whether or not a player hit many line drives for singles (which would result in a lower slugging percentage) or whether these line drives resulted in doubles, triples, or home runs (which would result in a higher slugging percentage). Thus, the model suggests that having a high line drive rate by itself does not necessarily result in higher slugging percentages. This is supported by the very low correlation between LDProp and slugging percentage (r=0.089).

Fly ball percentage, average exit velocity, speed score, and hard contact percentage are all significant predictor variables at the alpha=.05 level. Our highest VIF value among these variables is 2.29, implying that multicollinearity is not an issue with our baseline model.

It's worth noting at this point that we were interested to see if any significant interaction effects existed within our model; these include the interaction between hardProp and average exit velocity, fly ball Prop and average exit velocity, and line drive Prop and average exit velocity. It's reasonable to hypothesize that players who generally hit the ball hard (reflected in hardProp and LDProp) or in the air (reflected in FBProp) at a high exit velocity will achieve a higher slugging percentage. However, our testing resulted in no significant interaction effects. The results of this testing can be viewed in the appendix.

Our next step was to verify through stepwise regression that line drive percentage and soft contact percentage should be removed from the model, due to their insignificant p-values. We set both the alpha to remove and alpha to add at 0.10. The results are given in (Tables 5 and 6).

Table 5: ANOVA output of stepwise regression procedure predicting variation in slugging percentage.

Source	**DF**	**Adjusted SS**	**Adjusted MS**	**F-value**	**P-value**
Regression	4	0.76310	0.19077	88.01	0.000
FBProp	1	0.00843	0.00843	3.89	0.049
Speed	1	0.04028	0.04028	18.58	0.000
HardProp	1	0.15492	0.15492	71.47	0.000
Avg. Exit Velocity	1	0.11846	0.11846	54.65	0.000
Error	340	0.73702	0.00217		
Total	344	1.50011			
Model Summary:					
S	R^2	R^2 (adjusted)	R^2 (predicted)		
0.04656	50.87%	50.29%	49.42%		

Table 6: Model coefficient output from stepwise regression procedure predicting variation in slugging percentage

Term:	**Coefficient**	**Coefficient SE**	**T-Value**	**P-Value**	**VIF**
Constant	-0.636	0.105	-6.08	0.000	
FBProp	0.0780	0.0396	1.97	0.049	1.17
Speed	0.00677	0.00155	4.31	0.000	1.04
HardProp	0.5469	0.0647	8.45	0.000	1.76
Avg. Exit Velocity	0.00941	0.00127	7.39	0.000	1.55

The stepwise procedure confirms that FBProp, speed score, hardProp, and average exit velocity should be retained in the model as significant predictors, with softProp and LDProp being omitted from the model. Our R^2 value only drops from 0.5112 to 0.5087, meaning that these two variables did not significantly contribute to the explanation of variation within slugging percentage.

By individual variable test statistics, hardProp is the most significant term in the model (t=8.45), following by our variable of interest, average exit velocity (t=7.39). FBProp is the least significant term (t=1.97). Our highest variance inflation factor among the three variables is 1.76, implying that multicollinearity is not a concern for this model.

Our next step was to verify the assumptions of the ordinary least squares regression model. These assumptions are that the residuals of the model are normally distributed, with a mean of zero and a constant variance. Residual plots are given in (Fig. 4):

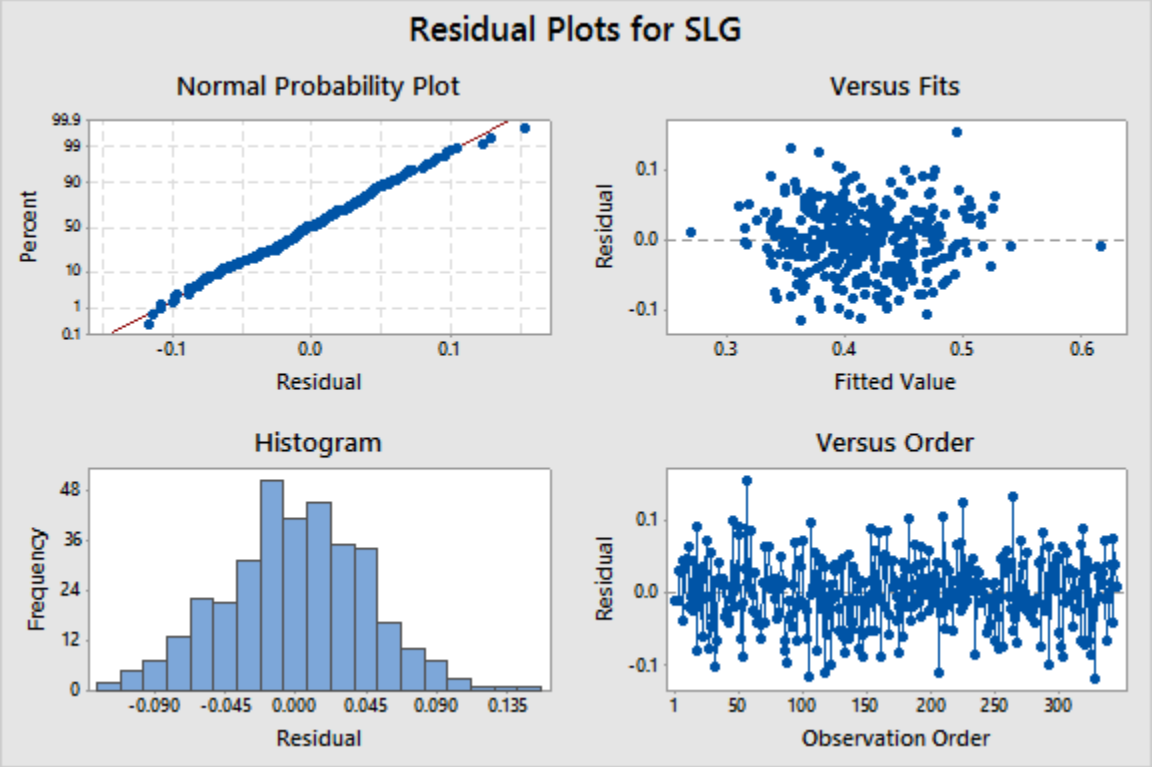

Figure 4. Residual plots from regression model generated by the stepwise procedure to predict variation in slugging percentage

In observing the residual plots, we see that the assumption of normally distributed residual terms with mean of zero and constant variance is upheld.

After we identified the significant independent variables in the model, our next step was to perform a best subsets procedure to see if any variables can be dropped without a significant reduction in explanation of the variance of slugging percentage, as simpler models are generally preferred over more complex models. If any variables are statistically significant but are practically insignificant in explaining the variation in slugging percentage, it would be ideal to remove them from the model. The results of the best subsets procedure are given in (Table 7).

Table 7: Best subsets output predicting variation in slugging percentage using FBProp, hardProp, speed, and average exit velocity as independent variables

Vars.	**R^2**	**R^2 (adj.)**	**R^2 (pred.)**	**Mall. CP**	**S**	**FBProp**	**HardProp**	**Avg. Exit Velo.**	**Speed**
1	40.2%	40.0%	39.4%	73.2	0.05112		X		
1	35.7%	35.5%	34.9%	104.0	0.05303			X	
2	47.8%	47.5%	46.9%	22.0	0.04783		X	X	
2	42.7%	42.4%	41.7%	57.3	0.05012		X		X
3	50.3%	49.9%	49.2%	6.9	0.04676		X	X	X
3	48.2%	47.7%	46.9%	21.6	0.04774	X	X	X	
4	50.9%	50.3%	49.4%	5.0	0.04656	X	X	X	X

Here, we see that the highest R^2 value for a one variable model is 0.402; 0.478 for a two variable model; 0.503 for a three variable model; and 0.509 for the model that includes all four variables. We notice that omitting FBProp only results in an R^2 reduction of .006, implying that

FBProp is the least significant variable in our model (which is supported by the fact that FBProp in our four variable model has a p-value of 0.049, which is statistically significant at alpha=0.05, but practically insignificant for our purposes). Because this is an insignificant drop-off, our final regression model will include hardProp, average exit velocity, and speed score as our explanatory variables. This is preferred over the three variable model that includes FBProp, hardProp, and average exit velocity, as this model has a lower R^2 value (.482 versus .503), higher Mallows CP value (21.6 versus 6.9) and higher regression standard deviation (.0477 versus .0468).

Our final step was to run the regression procedure including hardProp, speed, and average exit velocity as our independent variables and interpret the results (Tables 8 and 9).

Table 8: ANOVA output for final regression model predicting variation in slugging percentage

Source	**DF**	**Adjusted SS**	**Adjusted MS**	**F-Value**	**P-Value**
Regression	3	0.75466	0.25155	115.07	0.000
HardProp	1	0.20226	0.20226	92.52	0.000
Speed	1	0.03699	0.03699	16.92	0.000
Avg. Exit Velocity	1	0.11361	0.11361	51.97	0.000
Error	341	0.74545	0.74545		
Total	344	1.50011	0.00219		
Model Summary					
S	R^2	R^2 (adjusted)	R^2 (predicted)		
0.04676	50.31%	49.87%	49.18%		

Table 9: Coefficient output for final model predicting variation in slugging percentage

Term	Coefficient	Coefficient SE	T-Value	P-Value	VIF
Constant	-0.600	0.103	-5.80	0.000	
HardProp	0.5893	0.0613	9.62	0.000	1.56
Speed	0.00636	0.00155	4.11	0.000	1.03
Avg. Exit Velocity	0.00918	0.00127	7.21	0.000	1.53

We also need to ensure that the model assumption of normally distributed residuals with a mean of zero and a constant variance. We did this by examining the model's residual plots (Fig. 5).

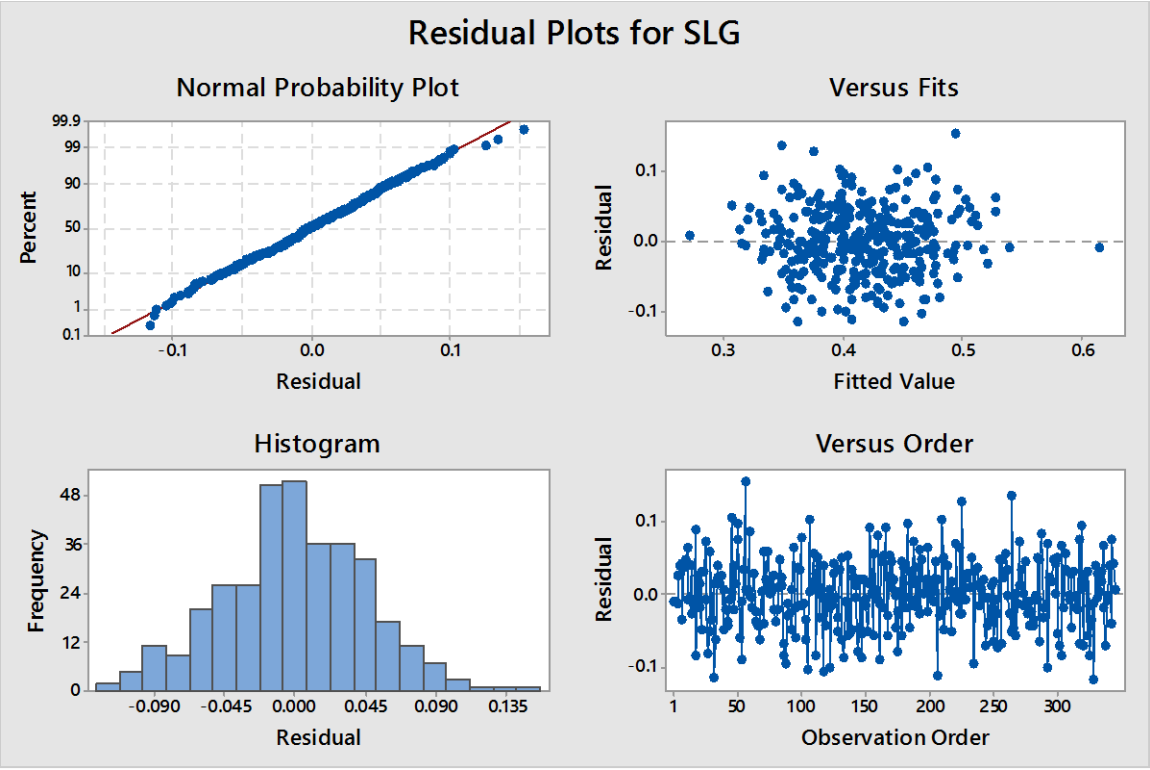

Figure 5: Residual plots from final regression model predicting variation in slugging percentage

Observing the residual plots, it's clear that the model assumption of normally distributed residuals with a mean of zero and constant variance is met.

Our final model is significant according to the overall F test (p-value < .0005), and each independent variable in the model is significant at the <.0005 level. According to the values of the individual t-test statistics, hard contact percentage is the most significant variable in the model (t=9.62), followed by average exit velocity (t=7.21), with speed score being the least significant (t=4.11).

The regression equation by our final model is given as:

$$\textit{Slugging percentage} = -0.600 + 0.5893(\textit{hardProp}) + 0.00636(\textit{spd}) + 0.00918(\textit{average exit velocity})$$

For each 0.01 increase in hardProp (which is equivalent to an increase of one percent), we expect a player's slugging percentage to increase by 0.5893*0.01=0.005893, or just under six points (a reminder to the reader that .001 represents one point of slugging percentage). For each one point increase in a player's speed score, we expect his slugging percentage to increase by .00636 (a reminder that a one point increase in speed score would be, as an example, an increase from four to five). And for each one MPH increase of a player's average exit velocity, we expect his slugging percentage to increase by .00918, or just over nine points. It's important to note that speed score operates on a scale from one to ten, while hardProp operates on a scale from 0 to 1. Average exit velocity theoretically has no upper bound, but practically, no player will have an average exit velocity above 100 MPH. In this case, we can say that average exit velocity can approximately range from 0 to 100 MPH. Thus, we can say that a one MPH increase in average exit velocity is scale equivalent to a 0.01 increase in hardProp, proportionally speaking. A one point increase in speed score is approximately equivalent to a ten point increase in hardProp or average exit velocity, proportionally speaking (due to the fact that speed score can range from 0 to 10). Because hardProp and average exit velocity are approximately scale equivalent, we can directly compare their coefficients. Thus, we see that, while the test statistic

of hardProp is more statistically significant than average exit velocity according to the t test statistic, according to their regression coefficients, average exit velocity carries a higher weight than hardProp, with a one MPH increase in average exit velocity being approximately 1.5 times as important as a one percent increase in hardProp (increase of 0.00918 for one MPH increase in average exit velocity versus an increase of 0.005893 for one percent increase in hardProp).

Explaining variation in BABIP

The second model we built attempts to explain the variation in BABIP using our independent variables as predictors. Because our dependent variable is BABIP, we first produced a correlation plot to examine the strength of the relationship between BABIP and average exit velocity, and we also examined the strength of the relationships between slugging percentage and each of the remaining independent variables in the model and compared these strengths to each other (Table 10).

Table 10: Correlation matrix between LDProp, FBProp, BABIP, softProp, hardProp, and average exit velocity.

	BABIP	**Avg. Exit Velocity**	**LDProp**	**FBProp**	**Speed**	**SoftProp**
Avg. Exit Velocity	0.188					
LDProp	0.418	0.007				
FBProp	-0.441	0.141	-0.297			
Speed	0.290	-0.086	-0.062	-0.153		
SoftProp	-0.177	-0.384	-0.355	-0.171	0.177	
HardProp	0.133	0.589	0.183	0.360	-0.164	-0.575

We notice that average exit velocity has a relatively minor association with BABIP, only yielding a correlation coefficient of 0.188. This tells us that merely striking the ball with a high speed on its own isn't a huge indicator of BABIP. The variable with the highest association is

FBProp (r=-0.441), followed by LDProp. This is unsurprising because line drives turn into hits very often, while fly balls fall in for hits infrequently.

We then ran the regression procedure to more thoroughly examine the relationship between the independent variables and BABIP, as well as to explain the variation in BABIP. We first developed the regression model with all predictor variables included (Tables 11 and 12).

Table 11: ANOVA output of full regression model predicting variation in BABIP

Source	**DF**	**Adjusted SS**	**Adjusted MS**	**F-Value**	**P-Value**
Regression	6	0.20778	0.03463	43.20	0.000
LDProp	1	0.02389	0.02389	29.80	0.000
FBProp	1	0.05426	0.05426	67.69	0.000
Speed	1	0.03898	0.03898	48.62	0.000
SoftProp	1	0.00085	0.00085	1.06	0.305
HardProp	1	0.00492	0.00492	6.14	0.014
Avg. Exit Velocity	1	0.00722	0.00722	9.00	0.003
Error	338	0.27094	0.00080		
Total	344	0.47873			
Model Summary					
S	R^2	R^2 (adjusted)	R^2 (predicted)		
0.02831	43.40%	42.40%	40.83%		

Table 12: Coefficient output of full regression model predicting variation in BABIP

Term	Coefficient	Coefficient SE	T-Value	P-Value	VIF
Constant	0.0514	0.0733	0.70	0.483	
LDProp	0.3171	0.0581	5.46	0.000	1.43
FBProp	-0.2190	0.0266	-8.23	0.000	1.43
Speed	0.00661	0.000947	6.97	0.000	1.05
SoftProp	-0.0540	0.0525	-1.03	0.305	1.71
HardProp	0.1113	0.0449	2.48	0.014	2.29
Avg. Exit Velocity	0.00238	0.00079	3.00	0.003	1.63

Our full regression model produces a coefficient of determination of 0.434, which means that just over forty-three percent of the variation in BABIP can be explained by the independent variables in the model. Our adjusted and predicted coefficients of determination are 0.424 and 0.4083 respectively, an insignificant reduction from our original R^2 value of 0.434. This tells us that we are not suffering a significant penalty for the number of independent variables in the model, nor does the model's accuracy decline when predicting new observations.

Examining the results of the coefficients table, we see that only softProp is insignificant at the alpha=0.05 level. Thus, softProp is not significantly different from mediumProp when taken into account with the other independent variables in the model.

LDProp, FBProp, hardProp, average exit velocity, and speed are all significant at the alpha=0.05 level. FBProp is the most significant predictor, with t=-8.23, while hardProp is the least significant of our remaining variables, with t=2.48. Our variable of interest, average exit velocity, is the second least significant variable with t=3.00 (p=0.003).

Our next step was to verify through stepwise regression that the softProp predictor should be removed from the model, due to its insignificant p-value. We set the alpha to remove and alpha to add both at 0.10 (Tables 13 and 14).

Table 13: ANOVA output of stepwise regression procedure predicting variation in BABIP

Source	**DF**	**Adjusted SS**	**Adjusted MS**	**F-Value**	**P-Value**
Regression	5	0.20694	0.04139	51.62	0.000
LDProp	1	0.03026	0.03026	37.74	0.000
FBProp	1	0.05346	0.05346	66.68	0.000
Speed	1	0.03828	0.03828	47.74	0.000
HardProp	1	0.00732	0.00732	9.12	0.003
Avg. Exit Velocity	1	0.00801	0.00801	9.99	0.002
Error	339	0.27179	0.27179		
Total	344	0.47873	0.00080		
Model Summary:					
S	R^2	R^2 (adjusted)	R^2 (predicted)		
0.02832	43.23%	42.39%	41.15%		

Table 14: Coefficient output of stepwise regression procedure predicting variation in BABIP

Term	Coefficient	Coefficient SE	T-Value	P-Value	VIF
Constant	0.0228	0.0678	0.34	0.737	
LDProp	0.3368	0.0548	6.14	0.000	1.28
FBProp	0.3368	0.0265	-8.17	0.000	1.42
Speed	-0.2163	0.00094	6.91	0.000	1.05
HardProp	0.00652	0.0421	3.02	0.003	2.01
Avg. Exit Velocity	0.00249	0.00079	3.16	0.002	1.60

The stepwise procedure confirms that FBProp, LDProp, speed score, hard contact percentage, and average exit velocity should be retained in the model as significant predictors, with softProp being omitted from the model. Our R^2 value only drops from 0.4340 to 0.4323, meaning that softProp does not significantly explain the variation in BABIP.

The results of the coefficient output show that our highest variance inflation factor is 2.01, indicating that multicollinearity is not a problem with our reduced model. By individual variable test statistics, FBProp is the most significant term in the model (t=-8.17), followed by speed (t=6.91), LDProp (t=6.14), average exit velocity (t=3.16), and hardProp (t=3.02). Thus, we see that our variable of interest, average exit velocity, is the second least significant variable in the model, but still featuring a very low p-value (p=0.002). This model shows that average exit velocity helps explain the variation in BABIP. More variables will be needed in future research, since R^2 is only 0.4323.

To ensure that our model meets the assumption of normally distributed residuals with a mean of zero and a constant variance, we examined the residual plots for this model (Fig. 6).

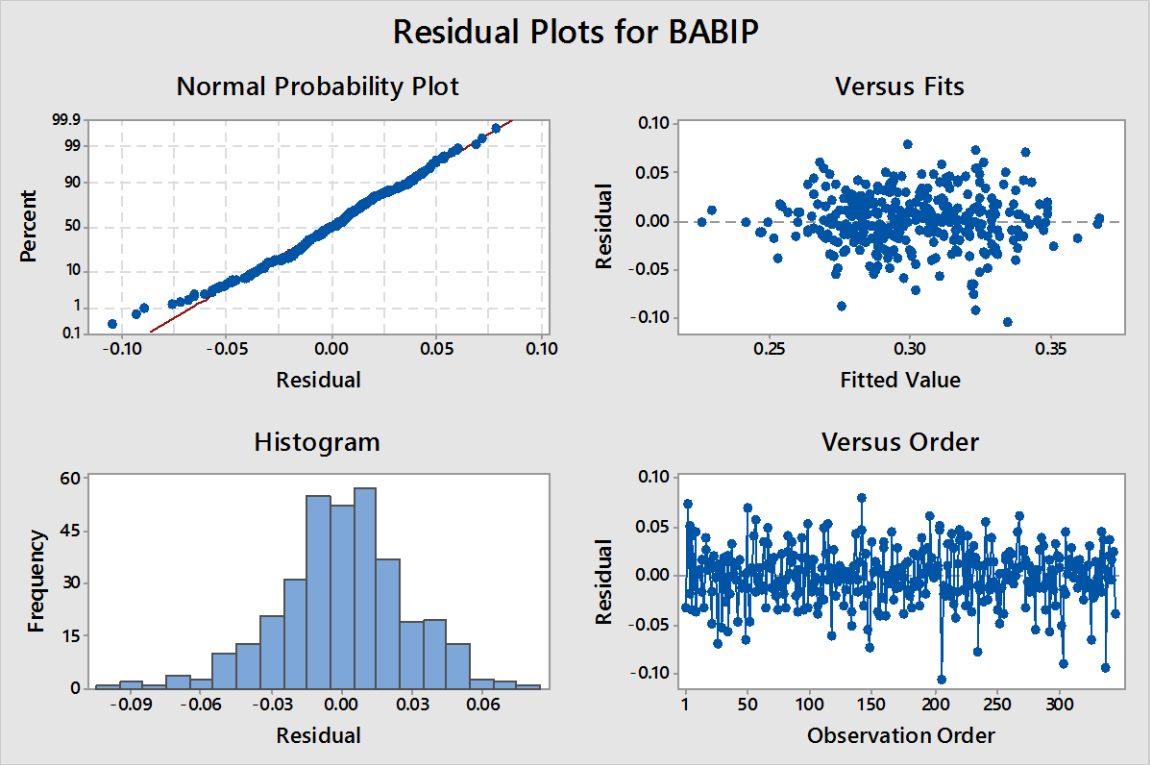

Figure 6. Residual plots from model generated by the stepwise procedure to predict variation in BABIP.

Observing the model residual plots, we see that the model assumptions are upheld. The residuals are approximately normal, with a mean of zero and a constant variance.

After we identified the statistically significant independent variables in the model, our next step was to perform a best subsets procedure to see if any variables can be dropped without a significant reduction in explanation of the variance of BABIP, as simpler models are generally preferred over more complex models. If any variables are statistically significant but are practically insignificant in explaining the variation in BABIP, it would be ideal to remove them from the model. The results of the best subsets procedure are found in (Table 15).

Table 15: Best subsets output predicting variation on BABIP using average exit velocity, FBProp, speed, LDProp, and hardProp as predictors.

Vars	R^2	R^2 (adj.)	R^2 (pred.)	Mall. CP	S	Avg. Exit Velocity	FBProp	Speed	LDProp	HardProp
1	19.4	19.2	18.5	140.0	0.03353		X			
1	17.5	17.2	16.5	151.8	0.03394				X	
2	29.2	28.8	27.8	83.6	0.03147		X			X
2	28.5	28.0	27.1	88.1	0.03164		X		X	
3	36.2	35.6	34.6	44.1	0.02993		X	X		X
3	35.3	34.7	33.7	49.5	0.03015		X	X	X	
4	41.7	41.0	39.9	13.1	0.02865	X	X	X	X	
4	41.6	40.9	39.8	14.0	0.02869		X	X	X	X
5	43.2	42.4	41.1	6.0	0.02832	X	X	X	X	X

We notice that a four variable model is preferred, as dropping down to a three variable model results in a significant reduction in R^2 of roughly six percent (0.417 versus 0.362), while using the full five variable model only results in an R^2 increase of about 1.5 percent (0.417 versus 0.432). The models that include four predictor variables have a coefficient of determination of just below 0.42, very close to our original model's R^2 value of 0.434. Both of the models with four indicator variables are practically identical in terms of R^2 values (0.417 versus 0.416), regression standard error (0.02865 versus 0.02869), and Mallows CP (13.1 versus 14). Because our primary research question involves using average exit velocity to explain the variation in BABIP, we will select the four variable model that includes exit velocity over the model that includes hardProp. However, this output indicates that average exit velocity is not nearly as

significant in explaining BABIP as it is for explaining slugging percentage, as in the BABIP model, the contributions of average exit velocity and hardProp are interchangeable. This is because the four variable model with average exit velocity and the four variable model with hardProp have about the same R^2 values, meaning they explain roughly the same amount of variation in BABIP.

Our final step was to run the regression procedure including FBProp, hardProp, speed, and average exit velocity as our independent variables and interpret the results (Tables 16 and 17).

Table 16: ANOVA output of final regression model predicting variation in BABIP.

Source	**DF**	**Adjusted SS**	**Adjusted MS**	**F-Value**	**P-Value**
Regression	4	0.19962	0.04991	60.79	0.000
FBProp	1	0.04639	0.04639	56.52	0.000
LDProp	1	0.04793	0.04793	58.38	0.000
Speed	1	0.03637	0.03637	44.30	0.000
Avg. Exit Velocity	1	0.03077	0.03077	37.48	0.000
Error	340	0.27911	0.00082		
Total	344	0.47873			
Model Summary					
S	**R^2**	**R^2 (adjusted)**	**R^2 (predicted)**		
0.02865	41.70%	41.01%	39.91%		

Table 17: Coefficient output of final regression model predicting variation in BABIP.

Term	Coefficient	SE Coefficient	T-Value	P-Value	VIF
Constant	-0.0881	0.0576	-1.53	0.127	
FBProp	-0.1818	0.0242	-7.52	0.000	1.16
LDProp	0.3960	0.0518	7.64	0.000	1.11
Speed	0.00634	0.00095	6.66	0.000	1.04
Avg. Exit Velocity	0.00391	0.00064	6.12	0.000	1.03

Our final model yields a coefficient of determination of .4170, with similar values for adjusted and predicted R^2. Thus, about 42 percent of the variation in BABIP can be explained by FBProp, LDProp, Speed, and average exit velocity, and we are not suffering a significant penalty by the number of variables included in the model, nor do we suffer a reduction in accuracy when predicting new values. Each predictor is significant at the p=.0005 level, and each VIF is 1.16 or lower, meaning that multicollinearity is not an issue with this model.

Our next step is to check to ensure that the model assumptions are satisfied (Fig. 7). Observing the residual plots, we see that the assumption of normally distributed residuals with a mean of zero and constant variance is upheld.

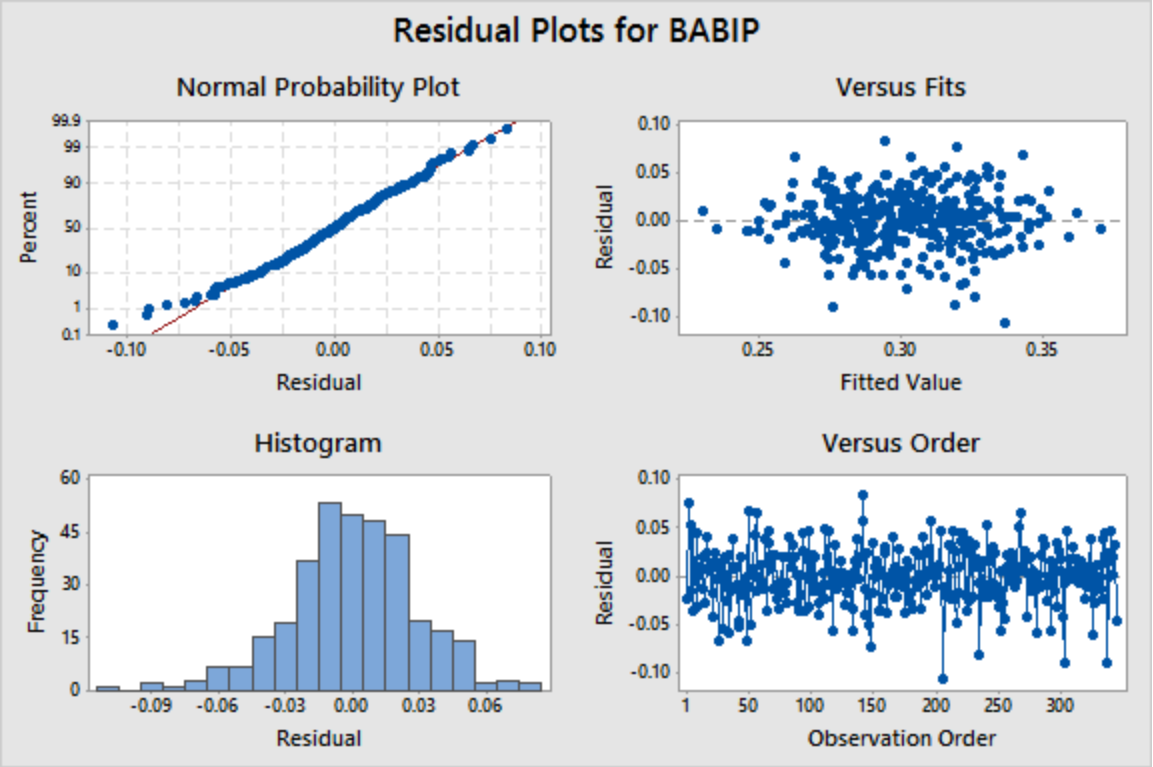

Figure 7. Residual plots from final regression model predicting variation in BABIP

The regression equation is given by:

BABIP = -0.0881 - 0.1818(FBProp) + 0.3960(LDProp) + 0.00634(Speed) + 0.00391(avg. exit velocity)

For each 0.01 increase in FBProp (which is equivalent to a one percent increase), we would expect BABIP to decrease by .1818*0.01=.001818, or roughly a two point decrease (a reminder to the reader that one point of BABIP is .001). For each 0.01 increase in LDProp (again, equivalent to a one percent increase), we would expect BABIP to increase by 0.01*.3960=.00396 about four points. For each one point increase in speed (a reminder that speed operates on a scale from 0 to 10), we expect BABIP to increase by .006344, about six points. And finally, for each MPH increase in average exit velocity, we expect BABIP to increase by .003911, or about four points. It is important to note that speed score operates on a scale from 0 to 10, while LDProp and FBProp operate on a scale from 0 to 1. Also, for the same

reasons stated in the previous model section, average exit velocity approximately operates on a scale from 0 to 100 MPH. Thus, proportionally speaking, a one point increase in speed is equivalent to an increase of 0.10 (or ten percent) in LDProp and FBProp, and a ten MPH increase in average exit velocity. So, for a 0.10 increase in FBProp, we expect BABIP to decrease by .01818, or about eighteen points. For a 0.10 increase (or ten percent) in LDProp we expect BABIP to increase by .0396, or just under forty points. And for a ten MPH increase in average exit velocity, we expect BABIP to increase by .0391, just under forty points. Thus, we can see that when we set these variables to be proportionally equivalent, speed actually has the lowest impact on a player's BABIP. LDProp and average exit velocity have the highest impact, with a ten point increase in these variables corresponding to a forty point increase in BABIP. This confirms our hypothesis that hitting the ball hard is the most important thing a player can do when it comes to having balls fall in for hits.

Model Validation

In order to see if the models obtained using 2015 season data will apply to future seasons, our final step in our analysis was to validate the models by using them to predict 2016 BABIP and slugging percentage, using 2016 player data. To do this, we generated a simple linear regression fitted line plot. The independent variable is the player's predicted 2016 slugging percentage or BABIP (given as xSLG and xBABIP), generated using the final regression equations obtained from the 2015 data, and the dependent variable is a player's actual 2016 slugging percentage or BABIP (given as SLG and BABIP). For slugging percentage, our final model independent variables are average exit velocity, speed score, and hardProp. For BABIP, our final model independent variables are FBProp, LDProp, average exit velocity, and speed score. We ended up selecting data on 371 players for our model validation.

The results for the model validation of slugging percentage and the fitted line plot are given in (Table 18) and (Fig. 8).

Table 18: ANOVA output of regression model predicting variation in actual slugging percentage from expected slugging percentage

Source	**DF**	**Sum of Squares**	**Mean Squares**	**F-Value**	**P-Value**
Regression	1	0.89468	0.89468	377.07	0.000
Error	369	0.87554	0.00237		
Total	370	1.77023			
Model Summary					
S	R^2	R^2 (adjusted)	R^2 (predicted)		
0.04871	50.54%	50.41%	50.00%		

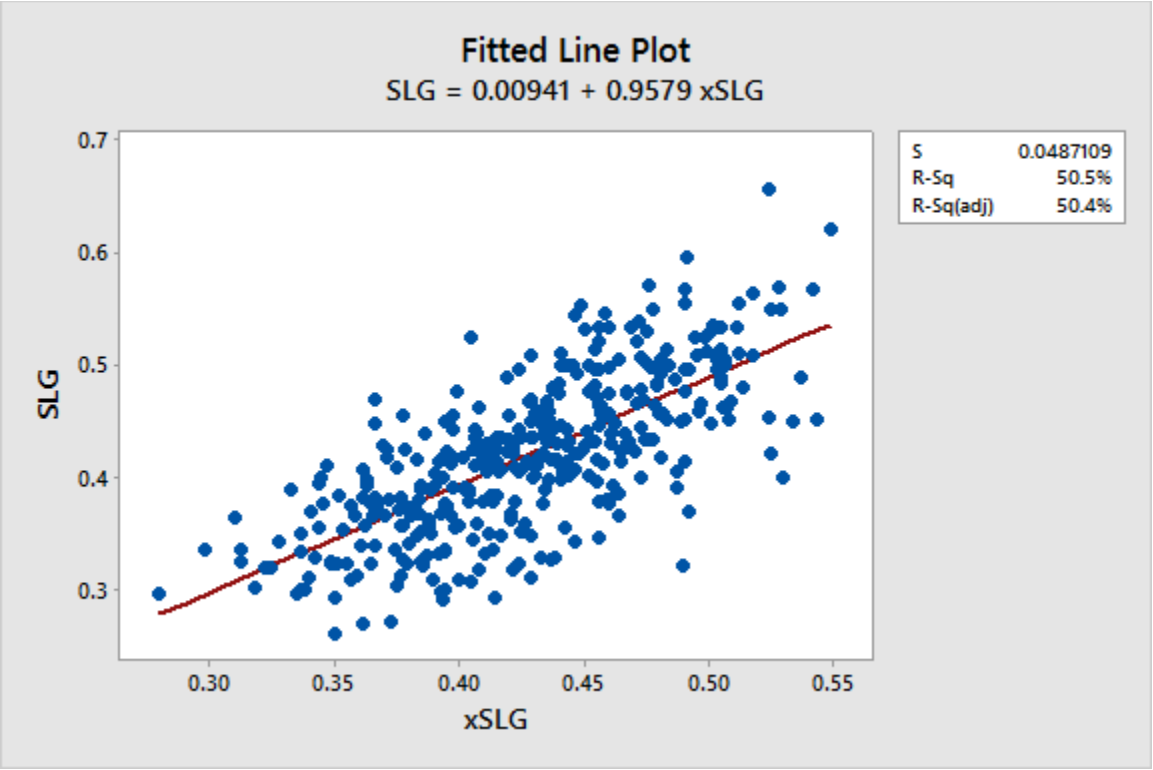

Figure 8. Fitted line plot and regression output for model predicting variation in actual slugging percentage from expected slugging percentage

The regression equation is given as:

$$SLG = 0.00941 + 0.9579 * xSLG$$

We can see that the simple linear regression model is significant ($p<.005$). Our R^2 value is given as 0.5054, compared to 0.5031 for our original 2015 model. The adjusted R^2 is 0.5041, compared to 0.4987 for the original model, which is actually a slight increase. The residual standard deviation of our validation model is 0.0487, compared to .0468 for the 2015 model. Thus, our validation model is virtually identical to our 2015 model.

Our final step in validating our slugging percentage model was to make sure the model assumptions are met. Residual plots for our validation model are given in (Fig 9).

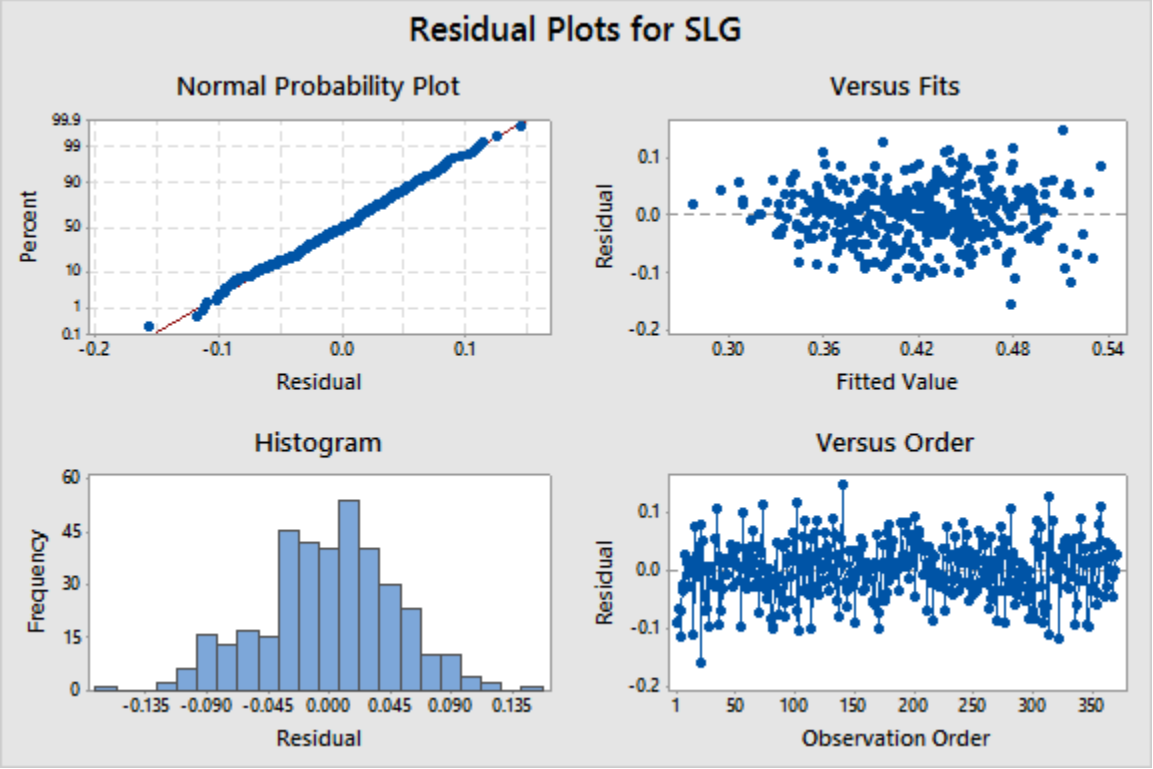

Figure 9. Residual plots from regression model predicting variation in actual slugging percentage from expected slugging percentage.

Observing the plots, we can see that the assumption of normally distributed residuals with a mean of zero and constant variance are met.

Because these results are very similar to that of our original model, we can conclude that, at least for the time being, our 2015 regression model predicting slugging percentage holds true regardless of season, barring any drastic future developments. We note that Statcast data has only been available for the 2015 and 2016 MLB seasons, and it's possible that the model's accuracy may shift as more seasons of Statcast data become available.

Next, we validated the model that explains the variation in BABIP. The ANOVA output and fitted line plot are found in (Table 19) and (Fig. 10).

Table 19: ANOVA output of regression model predicting variation in actual BABIP from expected BABIP

Source	**DF**	**Sum of Squares**	**Mean Squares**	**F-Value**	**P-Value**
Regression	1	0.19302	0.19302	243.53	0.000
Error	369	0.29246	0.00079		
Total	370	0.48547			
Model Summary					
S	R^2	R^2 (adjusted)	R^2 (predicted)		
0.02815	39.76%	39.60%	39.12%		

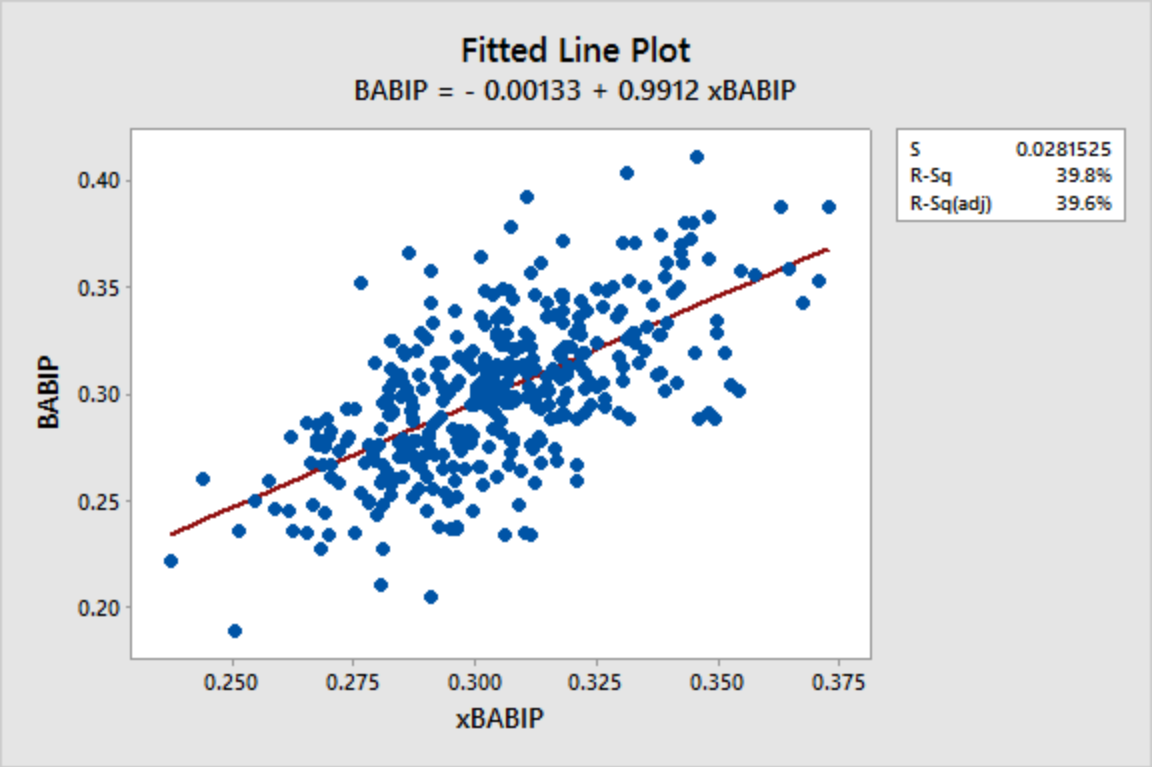

Figure 10. Fitted line plot and regression output for model predicting variation in actual BABIP from expected BABIP

The R^2 value of our validation model is given as 0.3976, compared to .4170 for our original model, a reduction of less than 0.02. Our adjusted R^2 is 0.3960, compared to 0.4101 in our original model. Our residual standard deviation is given as .0282, compared to .0287 for our original model. Thus, both models are virtually identical in terms of their diagnostics.

Our final step in validating our BABIP model was to make sure the model assumptions were met. Residual plots for our validation model are given in (Fig 11).

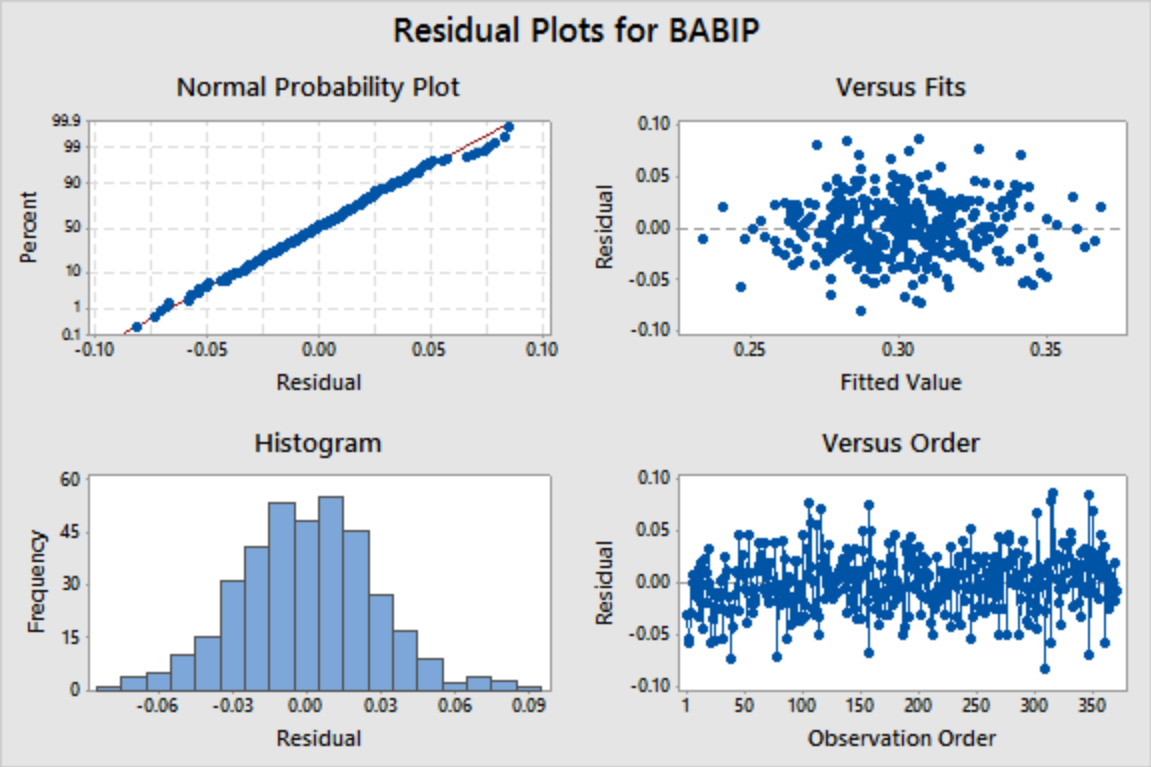

Figure 11. Residual plots of regression model predicting variation in actual BABIP from expected BABIP.

We can see that the model satisfies the assumption of normally distributed residuals with mean zero and constant variance.

Once again, our original 2015 model and our validation model are virtually identical, though there is a slightly larger reduction in R^2 compared to the slugging percentage validation model. We suspect this is due to BABIP being a much noisier statistic, as observed by comparing the fitted line plots for both variables. Thus, as we concluded with our slugging percentage model, our 2015 regression model should retain approximately the same level of accuracy regardless of season. Again, this is barring any significant fundamental changes moving forward as more and more Statcast data is collected across seasons.

CHAPTER FIVE: CONCLUSIONS

The deployment of Statcast introduced new data into the world of baseball analysis. Our goal in this research was to see if a player's average exit velocity, only made available in 2015, could significantly explain the variation in a player's BABIP and slugging percentage when taken into account with several other traditional baseball statistics. Through producing regression models, we discovered that average exit velocity can significantly explain the variation in a player's slugging percentage. Our model was able to explain about half the variation in slugging percentage. Average exit velocity ended up being one of the most significant explanatory variables in our model, with each increase of one MPH leading to a nearly ten point increase in a player's slugging percentage. This ended up being about one and a half times more significant than a one point increase in hardProp. This confirms what we intuitively hypothesized – players who strike the ball on average with a high velocity tend to hit for more power than players who don't. The applications of this knowledge are immediately clear: general managers who are looking to add power to their lineup should take into account a free agent's average exit velocity when looking at their statistical profile. We expect that a player's average exit velocity is a good true talent indicator, as it's difficult to believe that a player could consistently make hard contact through sheer luck.

When we produced our regression model to explain the variation in BABIP, we could only explain about forty percent of the variation in BABIP when we included average exit velocity, LDProp, FBProp, and speed in the model – about ten percent lower than our coefficient of determination in the slugging percentage model. We suspect this is due to the fact that BABIP is a much noisier statistic that is influenced by a bevy of factors outside our model that are difficult to quantify, such as the ballpark being played in, the quality of the defense, and the ability of the opposing pitcher to limit hard contact. Nevertheless, we're encouraged by the observation that average exit velocity has one of the largest weights in our regression model according to our regression equation, despite it not being the most statistically significant

variable in the model. This confirms that average exit velocity is an important factor in explaining BABIP, when taken into consideration with other variables.

Future research into Statcast data could consider the impact that average exit velocity has on variables besides slugging percentage and BABIP, as well as considering launch angle off the bat, something we did not look at with this study. Making solid contact involves more than hitting the ball hard; it also involves striking the ball with an ideal launch angle, one that more often results in line drives rather than ground balls or fly balls. Also, as more seasons of Statcast data are collected, the conclusions of this study could be tested to see if they apply to future seasons, or if the 2015 and 2016 seasons end up being outliers from future trends.

WORKS CITED

Arthur, Rob. "Chase Utley is the Unluckiest Man in Baseball." *FiveThirtyEight*. ESPN, 15 May 2015. Web. 15 March 2017.

Berg, Ted. "MLB's new Statcast technology will change the way you watch baseball." *USA Today Sports*. USA Today, 6 May 2015. Web. 15 March 2017.

Birnbaum, Phil. "A Guide to Sabermetric Research." *SABR*. Society for American Baseball Research, n.d. Web. 5 March 2017.

Casella, Paul. "Statcast primer: Baseball will never be the same." *MLB.com*. Major League Baseball, 24 Apr. 2015. Web. 5 March 2017.

"Complete List (Offense)." *Fangraphs*. Fangraphs, n.d. Web. 27 March 2017

"Exit Velocity (EV)." *MLB.com*. Major League Baseball, n.d. Web. 27 March 2017

Fangraphs. Fangraphs, n.d. Web. 20 March 2017

Frost, Jim. "Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables." *The Mintab Blog*. Minitab, 13 June 2013. Web. 27 Mar. 2017.

"Multicollinearity in regression." *Minitab Support*. Minitab, n.d. Web. 27 March 2017.

Petriello, Mike. "3 cool lessons from Statcast's debut season." *MLB.com*. Major League Baseball, 12 Nov. 2015. Web. 1 March 2017

Willman, Daren. "About." *Baseball Savant*. N.p., n.d. Web. 27 March 2017.

Shaw, Stephen. "Updated MLB Statcast Data (July 2015)." *Banished To The Bullpen*. Wordpress, 1 July 2015. Web. 15 March 2017.

Simon, Eric. "Sabermetrics And You: The Big Three, Part 1 - Batting Average." *Amazin' Avenue*. SB Nation, 09 Dec. 2010. Web. 8 Apr. 2017.

Stampel, Billy. "Barrels, Normative Analysis, and the Beauties of Statcast." *The Hardball Times*. The Hardball Times, 29 Sept. 2016. Web. 15 March 2017.

APPENDIX

Table A1. Summary of baseball statistics used in this thesis ("Complete List…") and ("Exit Velocity…")

Statistic	**Summary**
Batting average on balls in play (BABIP)	The rate at which the batter gets a hit when he puts the ball in play, calculated as (H-HR)/(AB-K-HR+SF).
Slugging percentage (SLG%)	Average number of total bases per at bat, calculated as Total Bases/AB.
Line drive proportion (LDProp)	The proportion of a batter's balls in play that are line drives, calculated as LD/BIP.
Fly ball proportion (FBProp)	The proportion of a batter's balls in play that are fly balls, calculated as FB/BIP.
Soft contact proportion (SoftProp)	Proportion of a batter's soft-hit batted balls.
Hard contact proportion (HardProp)	Proportion of a batter's hard-hit batted balls.
Speed score	A statistic that attempts to measure a player's running speed and ability. Ranges from zero to ten.
Average exit velocity	The average velocity, in MPH, of the ball when a batter puts the ball in play.

Table A2. ANOVA output of full regression model predicting variation in slugging percentage with interaction terms

Source	**DF**	**Adjusted SS**	**Adjusted MS**	**F-Value**	**P-Value**
Regression	9	0.77635	0.08626	39.93	0.000
Speed	1	0.03795	0.03795	17.57	0.000
SoftProp	1	0.00003	0.0000	0.02	0.899
HardProp	1	0.00142	0.00142	0.66	0.419
Avg. Exit Velocity	1	0.00000	0.00000	0.00	0.962
LDProp	1	0.00004	0.00004	0.02	0.890
FBProp	1	0.00349	0.00349	1.62	0.204
FBProp*Avg. Exit Velocity	1	0.00457	0.00457	2.11	0.147
HardProp*Avg. Exit Velocity	1	0.00285	0.00285	1.32	0.252
LDProp*Avg. Exit Velocity	1	0.00007	0.00007	0.03	0.855
Error	335	0.72376	0.00216		
Total	344	1.50011			
Model Summary					
S	**R^2**	**R^2 (adjusted)**	**R^2 (predicted)**		
0.04648	51.75%	50.46%	48.72%		

Table A3. Coefficient output of full regression model predicting variation in slugging percentage with interaction terms

Term	Coefficient	SE Coefficient	T-Value	P-Value	VIF
Constant	0.210	0.696	0.30	0.763	
Speed	0.00654	0.00156	4.19	0.000	1.06
SoftProp	0.0111	0.0869	0.13	0.899	1.74
HardProp	-1.19	1.47	-0.81	0.419	905.72
Avg. Exit Velocity	-0.00037	0.00781	-0.05	0.962	58.38
LDProp	-0.39	2.80	-0.14	0.890	1234.52
FBProp	-0.696	0.547	-1.27	0.204	225.12
FBProp*Avg. Exit Velocity	0.00865	0.00595	1.45	0.147	241.65
HardProp*Avg. Exit Velocity	0.0190	0.0165	1.15	0.252	1116.61
LDProp*Avg. Exit Velocity	0.0058	0.0316	0.18	0.855	1280.44