STA363 Final Project

Daniel Beumer

2024-04-26

Contents

Abstract

A common metric used to evaluate the performance of a pitcher in baseball is wOBA, or weighted on-base average. This report analyzes three different methods of predicting wOBA and how these methods provide value to a baseball organization that wants to use wOBA to drive its decisions. The data are comprised of many different metrics recorded during all games of the 2023 MLB season, with these metrics representing manipulations of counting statistics as well as ball flight data gathered by technology like Trackman. We discuss how to predict wOBA and what features from the data are the most influential on wOBA by utilizing K-Nearest Neighbors, elastic net, and random forests. The outcome of our analysis was that elastic net proves to be the best at predicting wOBA, while random forests also provide value in scouting pitchers when our data may be limited. We believe that the use of these methods as compliments and in the right circumstances will drive better decisions as an organization in terms of pitcher scouting and acquisition.

Section 1: Introduction

Our goal is to determine how well this data can be used to predict wOBA, as this will allow us to figure out what pitchers to approach via trades or free agency. This metric is, in our opinion, the best way of evaluating a pitcher's overall value given the data set provided because it balances how hitters perform against the pitcher. This is important because all teams in the MLB are also faced with this broad question, and we need to be the best at predicting a pitcher's value so that we can make better decisions than our opposition in terms of acquiring these pitchers.

Our data has 647 observations and 19 columns before cleaning. This breaks down to 647 pitchers and roughly 16 statistics that we are considering. Some notable features include: the amount of batters faced by a pitcher in 2023, the percent of hitters that a pitcher walked in 2023, the percent of hitters that a pitcher struck out in 2023, the average exit velocity on balls hit off the pitcher (in miles per hour) in 2023, the percent of batted balls considered "hard hit" balls off the pitcher in 2023, and the percent of pitches thrown in the strike zone by a pitcher in 2023.

This link leads to the exact data set pulled from Baseball Savant, and a glossary is provided at the bottom of this paper that describes each of our features.

Section 2: Data Cleaning and EDA

Data Cleaning

Since our data comes from a reliable source that tracks baseball well, we do not need to alter our data much. We only have to perform two tasks.

The first thing we have to do is remove columns that are not useful to us, which in this case are the columns "last_name, first_name", "player_id", "year", and "avg_best_speed". These are simply identifiers that are used to show viewers of the site what player these stats belong to and the year. In our case, we do not need to identify players, as we are solely interested in how these statistics affect wOBA. Additionally, these are all statistics from the same year, so the column representing year is completely useless. Finally, we are going to remove the average of the softest 50% of batted balls allowed, represented as "avg_best_speed". We are doing this because we already have average exit velocity, and this measures the same thing from a larger and more "interpretable" set of the pitcher's exit velocities given up.

The second alteration we are making to the data is to convert our percentages to scale from 0 to 1 instead of 0 to 100. This is because our response variable, and some other explanatory variables, have a similar scale, and thus this will put their coefficients at similar levels to the other variables. If this is not performed, for example, the coefficient for BABIP is much higher than the coefficient for percent variables, as it's measured on the same scale as wOBA while the others are much higher. Thus, this change is made in order to make the coefficients easier to interpret, though it is not fully necessary.

Our final data set for use has 647 observations of 15 variables, or 15 different statistics recorded for 647 different pitchers in 2023.

Exploratory Data Analysis

We must first fit a linear model so that we can check the conditions for regression. The coefficients of this model are displayed in Figure 2.1. These give us an idea of what has influence on wOBA, but these should not be taken at face value. We must first assess the conditions for regression. It is also important to note that features like plate appearances or average launch angle will have smaller coefficients regardless of their impact because they have a much larger scale than wOBA.

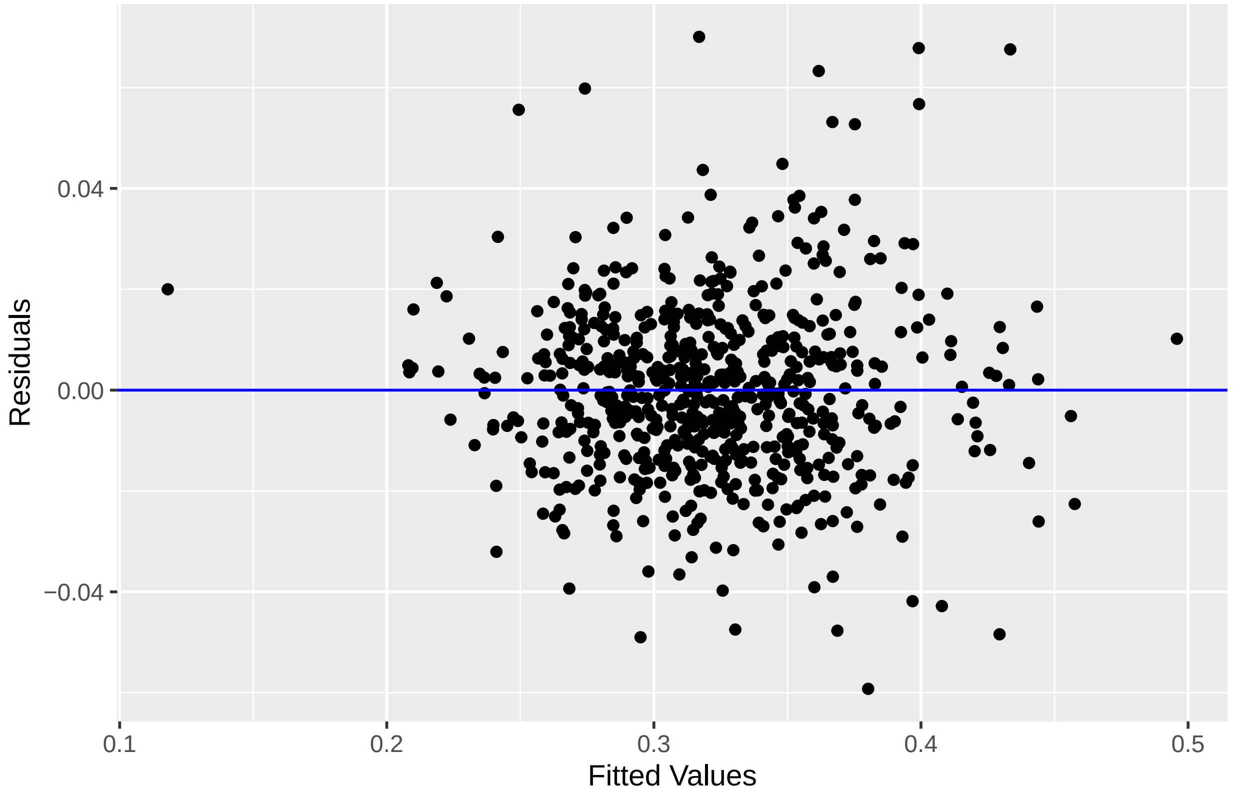
Table 1: Figure 2.1: Linear Model Coefficients

	x
(Intercept)	0.1840642
pa	-0.0000070
k_percent	-0.4022401
bb_percent	0.2833113
babip	0.5739703
exit_velocity_avg	-0.0006132
launch_angle_avg	0.0006769
sweet_spot_percent	0.0255056
barrel_batted_rate	0.6002211
hard_hit_percent	0.0297009
oz_swing_percent	0.0067100
in_zone_percent	0.0157880
whiff_percent	0.0248232
swing_percent	-0.0606253
pull_percent	0.0621587

Zero Mean & Constant Variance

In order to assess if the residuals have constant variance and zero mean, we can examine Figure 2.2 below.

Figure 2.2: Residuals Plot

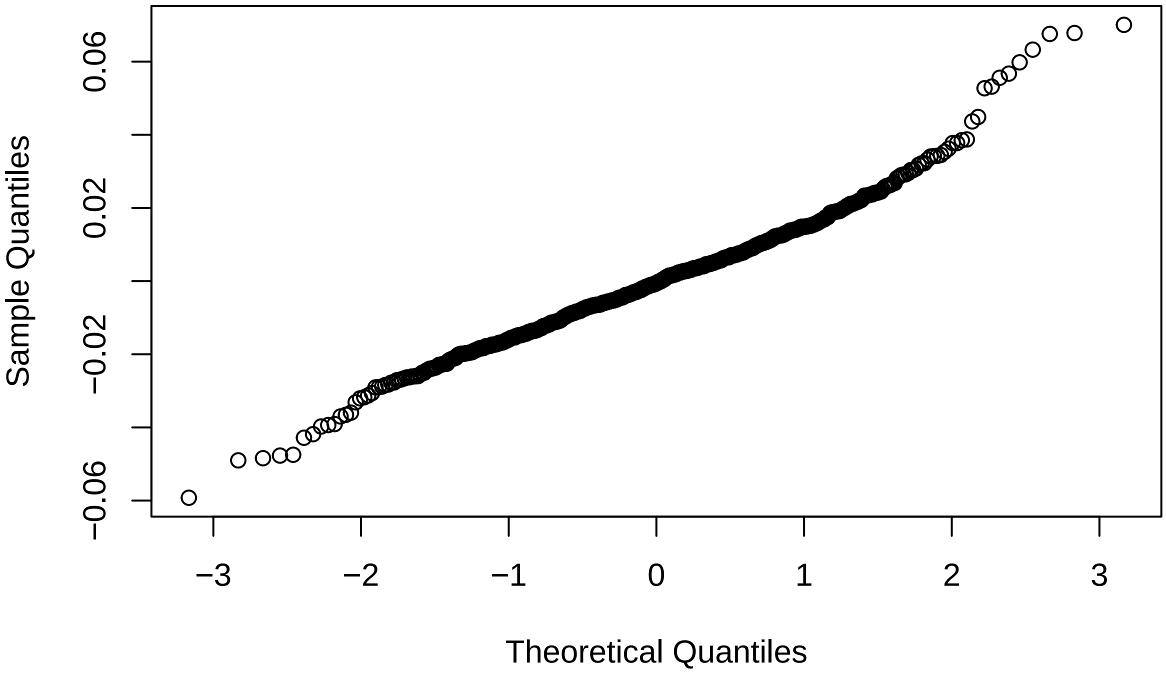


The zero mean condition appears to be satisfied, as the mean of the residuals in this graph is zero. Additionally, the residuals have constant variance for the most part. The variance does increase as the fitted values deviate from the middle, however, I believe the constant variance condition is met.

Normality

The residuals appear to satisfy the normality assumption, as the QQ Plot in Figure 2.3 shows a linear relationship.

Figure 2.3: Residuals QQ Plot



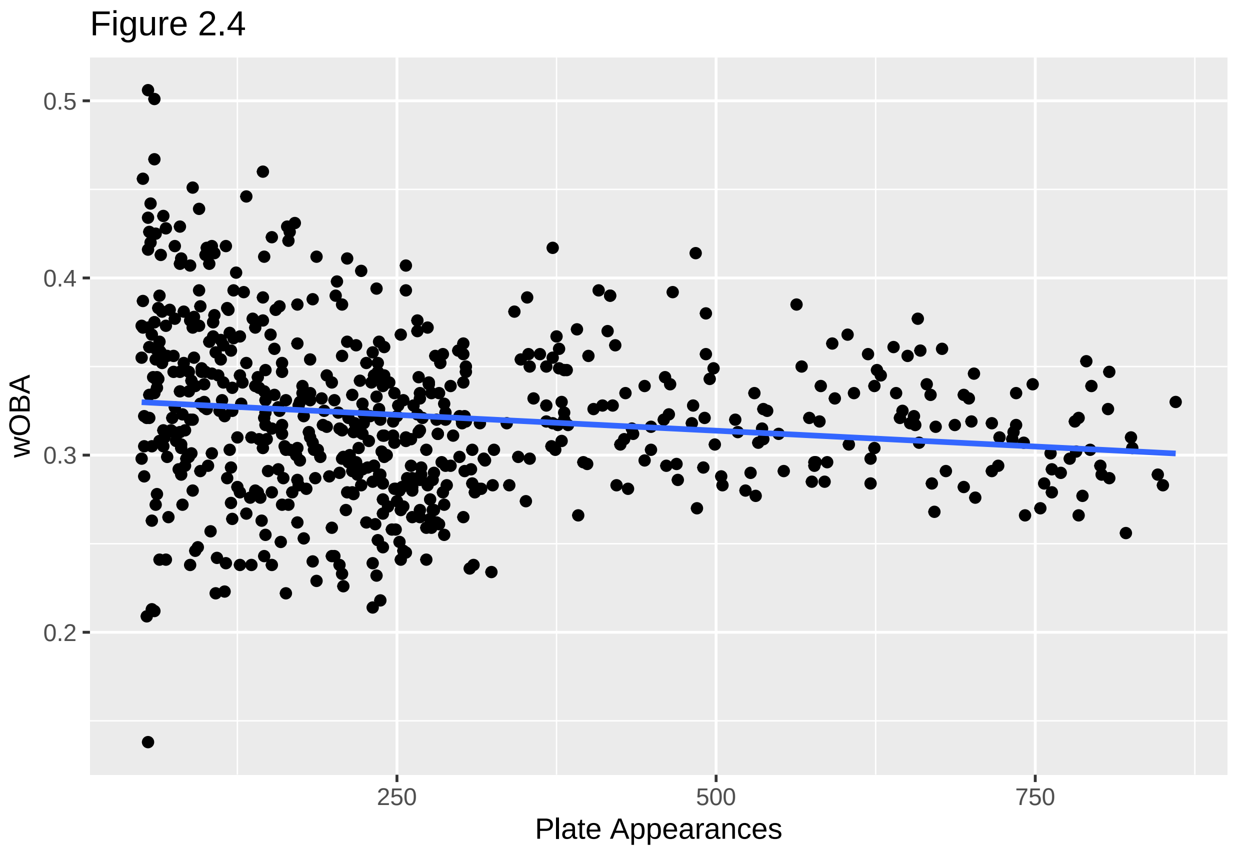
Independence

We assume the residuals are independent, given that we know how these statistics are recorded and formulated as well as what they represent.

Linearity

The linearity condition will be checked by examining each explanatory variable one by one to confirm that they all have a linear relationship with wOBA. Since we are only using numerical variables, we will examine a scatter plot with the explanatory variable on the x-axis and wOBA on the y-axis with the hopes of seeing a linear relationship displayed.

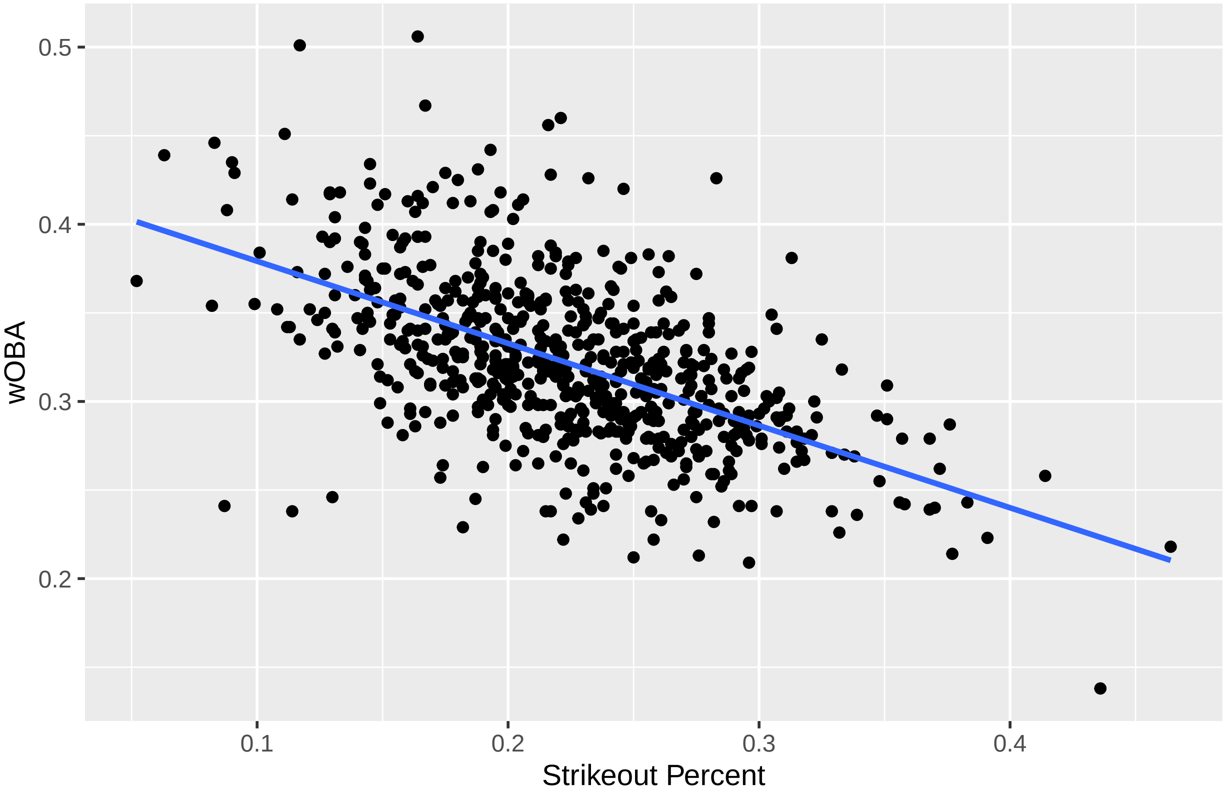
Plate Appearances:



According to Figure 2.4, there appears to be a negative linear relationship between plate appearances and wOBA. There are much more pitchers with lower plate appearances, thus causing the wOBA to vary more for those numbers, but on average there does appear to be a linear relationship between the two variables. This makes sense to a degree, as pitchers who perform well (thus have lower wOBA) and are not injured will likely be used the most by teams. They are less susceptible to being removed from the roster due to poor performance.

Strikeout Percent:

Figure 2.5



By Figure 2.5, there appears to be a negative linear relationship between wOBA and strikeout percent. This is expected, as high strikeout percentage is generally associated with better performance.

Walk Percent:

Figure 2.6

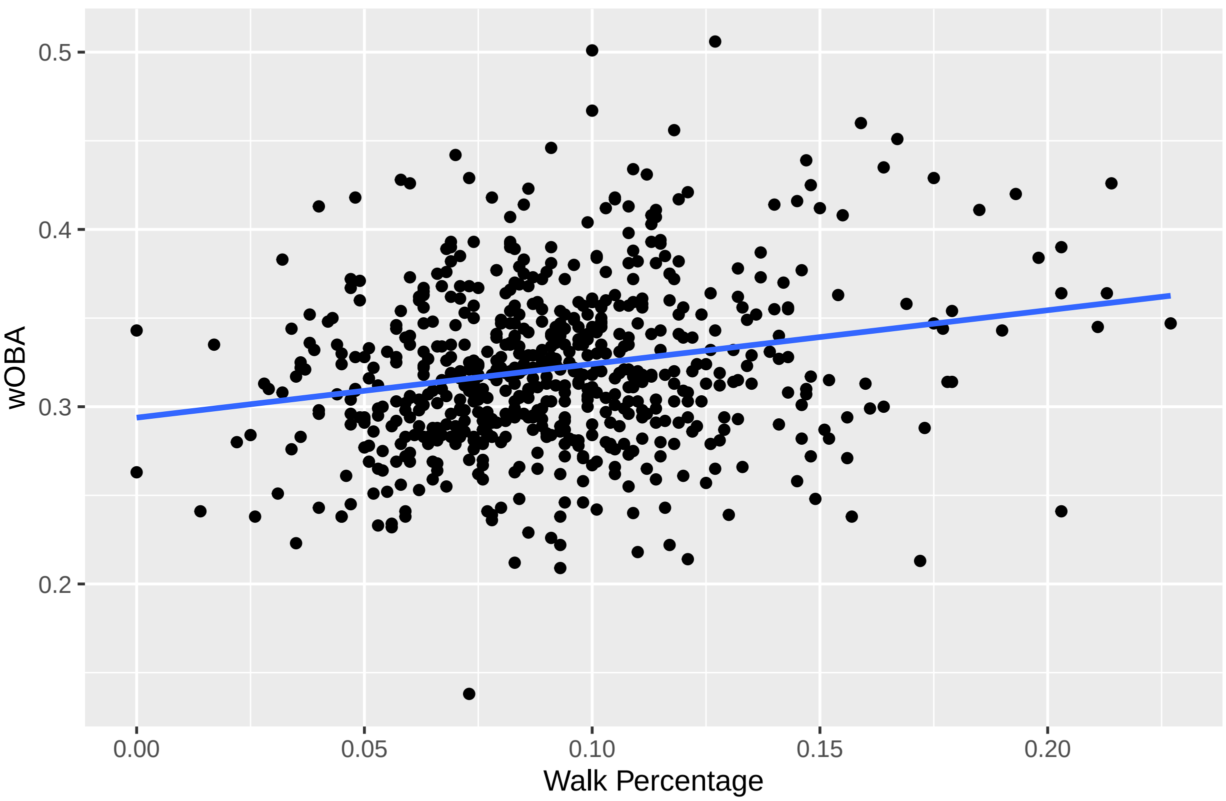


Figure 2.6 shows a positive linear relationship between walk percent and wOBA. This is expected as well, as walks are generally associated with poorer performance.

BABIP (Batting Average on Balls in Play):

Figure 2.7

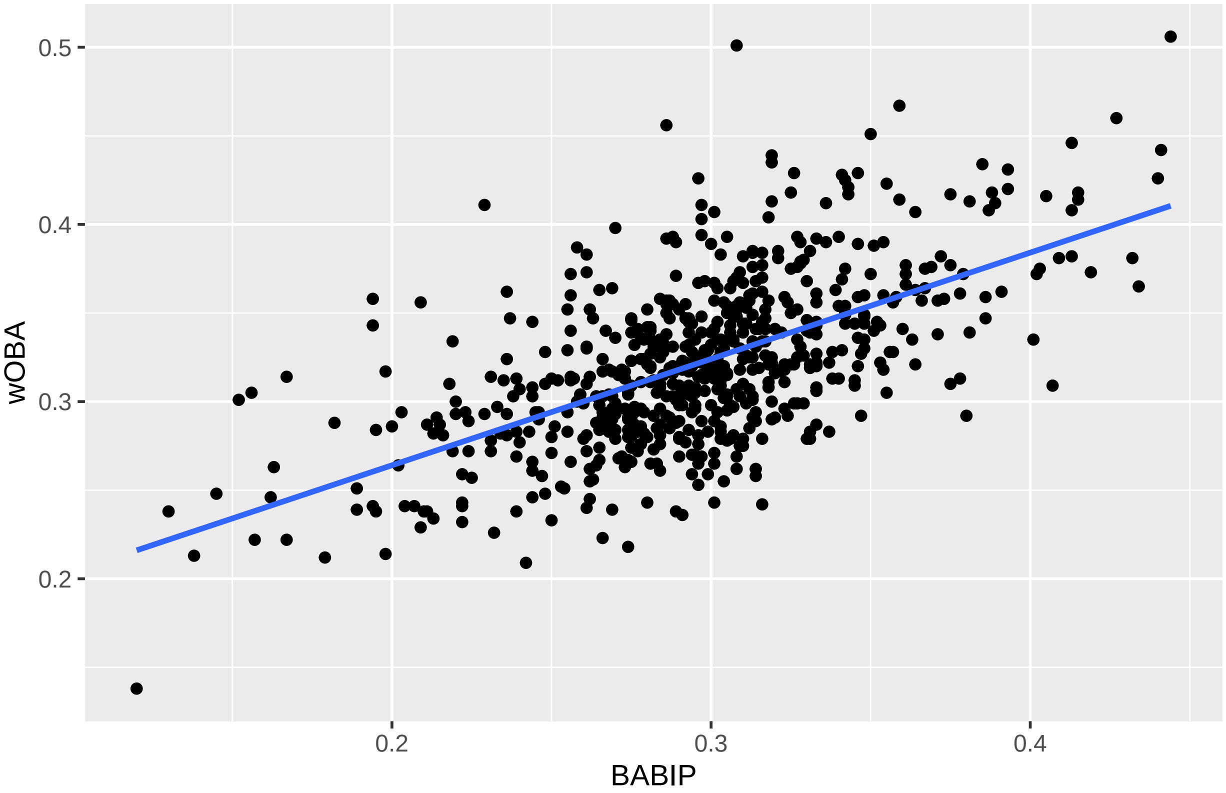
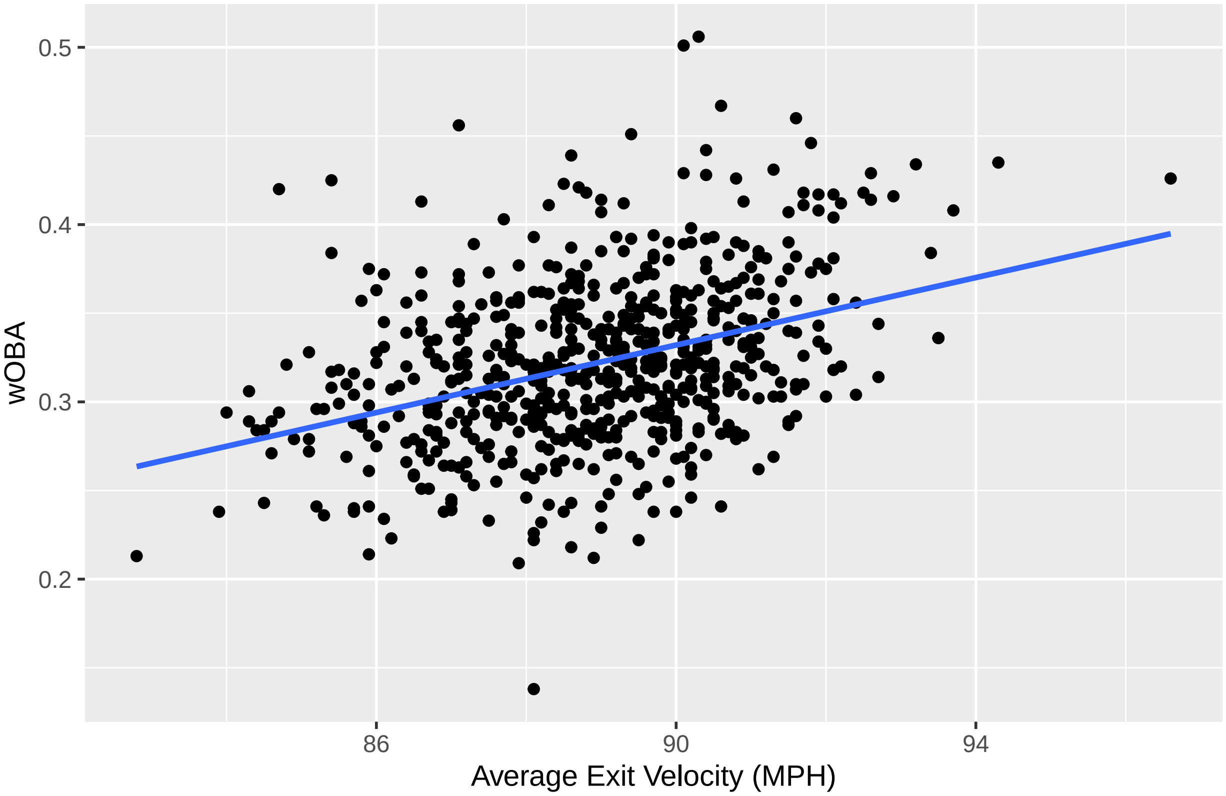


Figure 2.7 shows a clear positive linear relationship between BABIP and wOBA. This also makes intuitive sense. As the percentage of balls put in play resulting in hits increases, it would follow that a pitcher's performance decreases.

Average Exit Velocity:

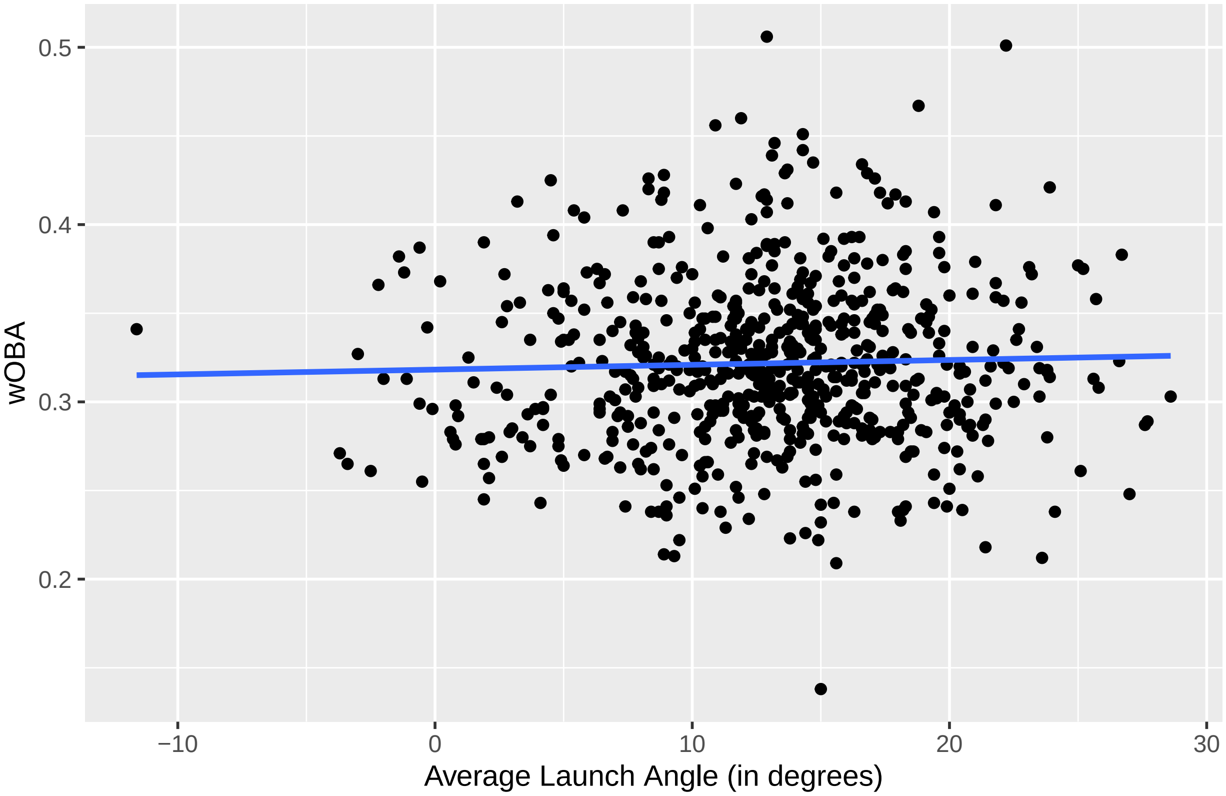
Figure 2.8



There appears to be a positive linear relationship between average exit velocity and wOBA per Figure 2.8. Once again, this is expected. Pitchers that allow harder hit balls on average should have worse performance metrics.

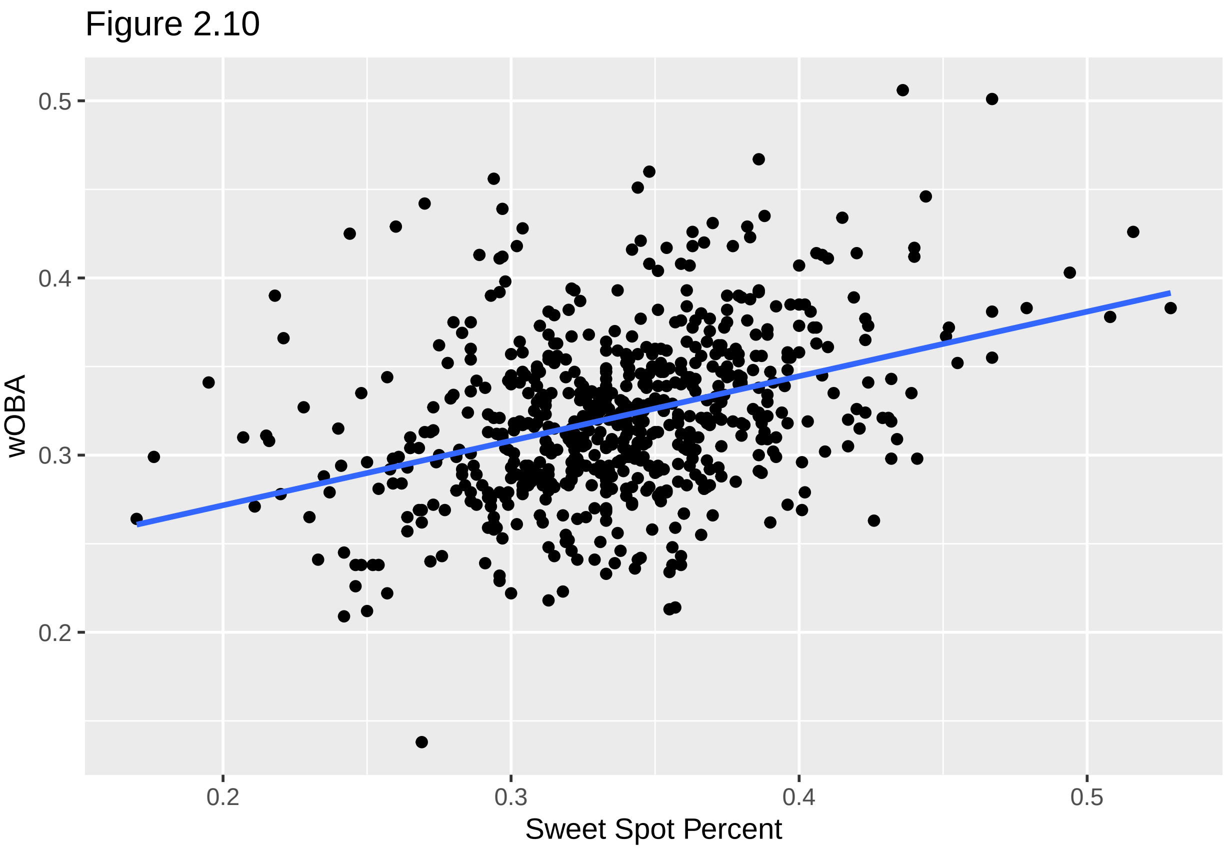
Average Launch Angle:

Figure 2.9



Considering the scale of launch angle is greater than that of wOBA, there does appear to be a slight relationship between wOBA and average launch angle according to Figure 2.9.

Sweet Spot Percent:



There appears to be a positive linear relationship between sweet spot percent and wOBA, as seen in Figure 2.10. Considering a higher sweet spot percent indicates a pitcher is giving up better hit balls more often, this result is expected.

Barrel Rate:

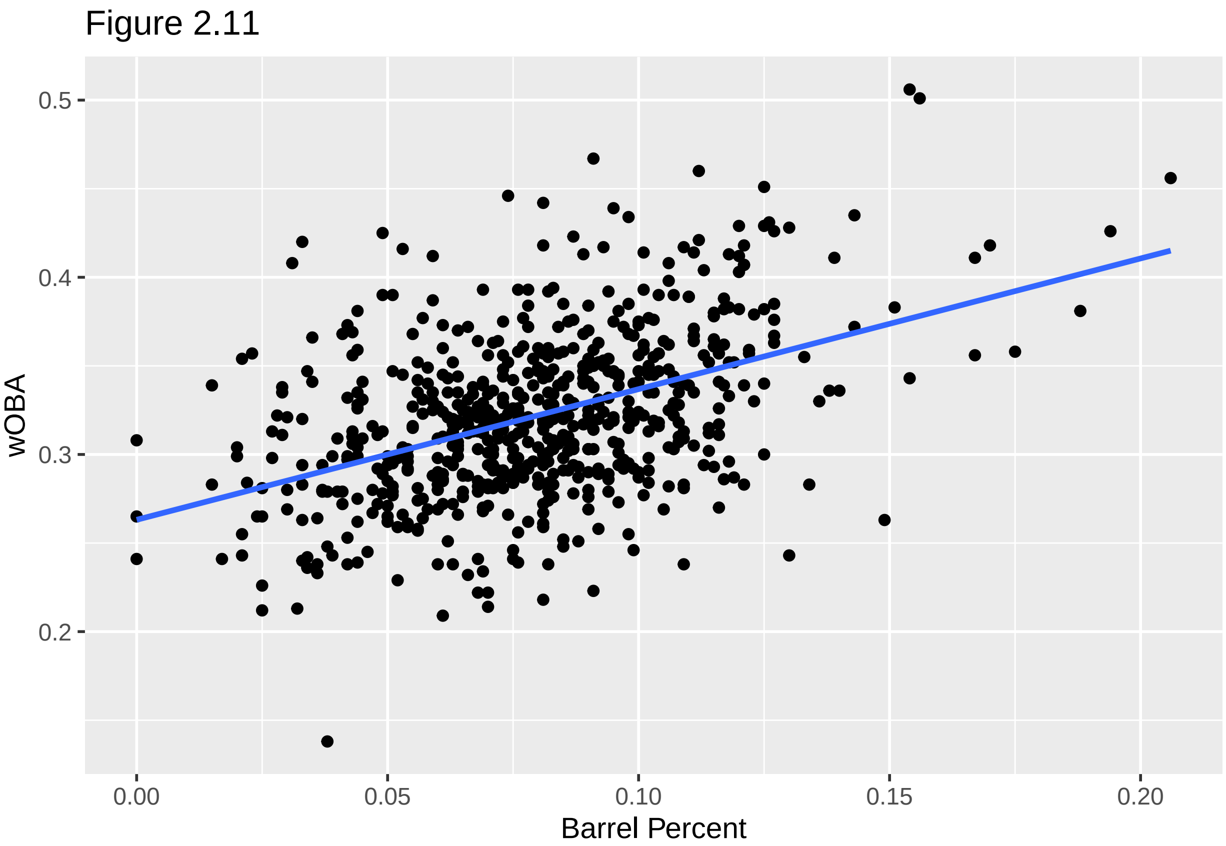
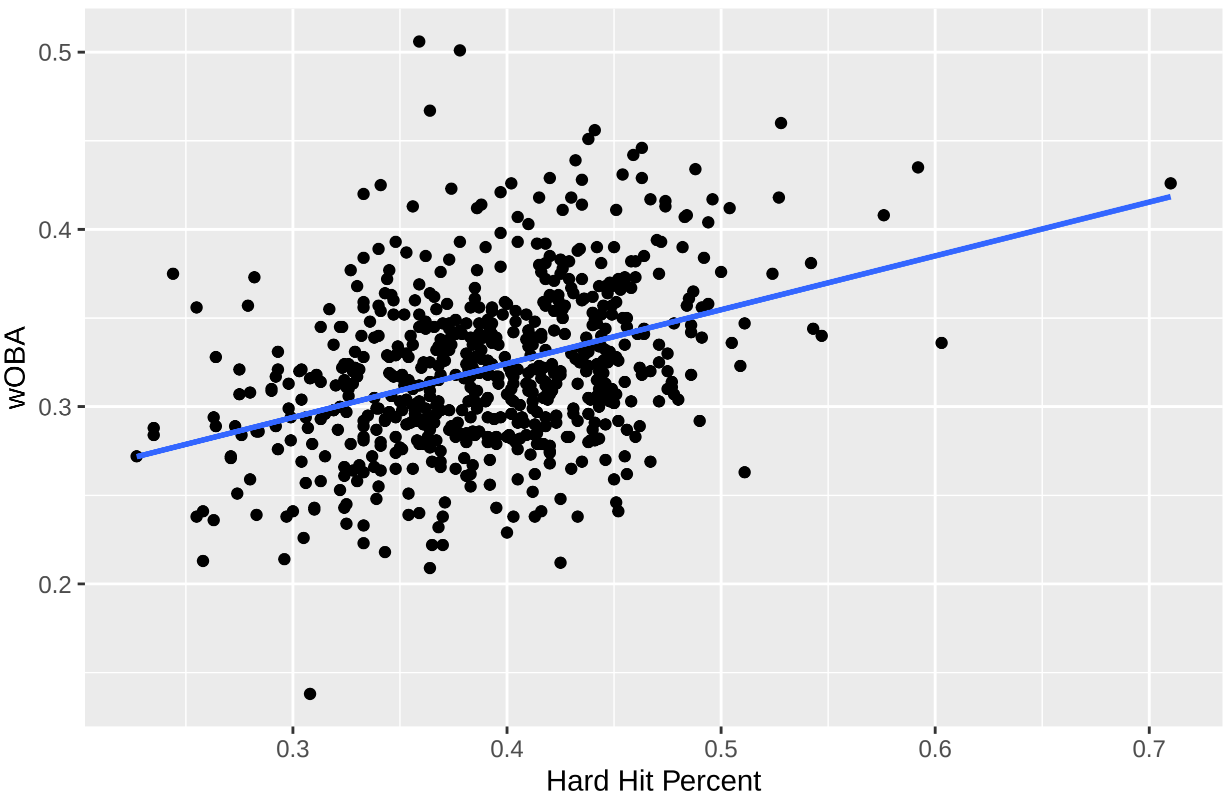


Figure 2.11 shows a positive linear relationship between barrel rate and wOBA. This follows as expected for the same reasons as sweet spot percent.

Hard Hit Percent:

Figure 2.12



The relationship between hard hit percent and wOBA is positive and linear, as displayed in Figure 2.12. As a greater allowed hard hit percent indicates poorer pitcher performance, this is expected.

Out-of-Zone Swing Percent:

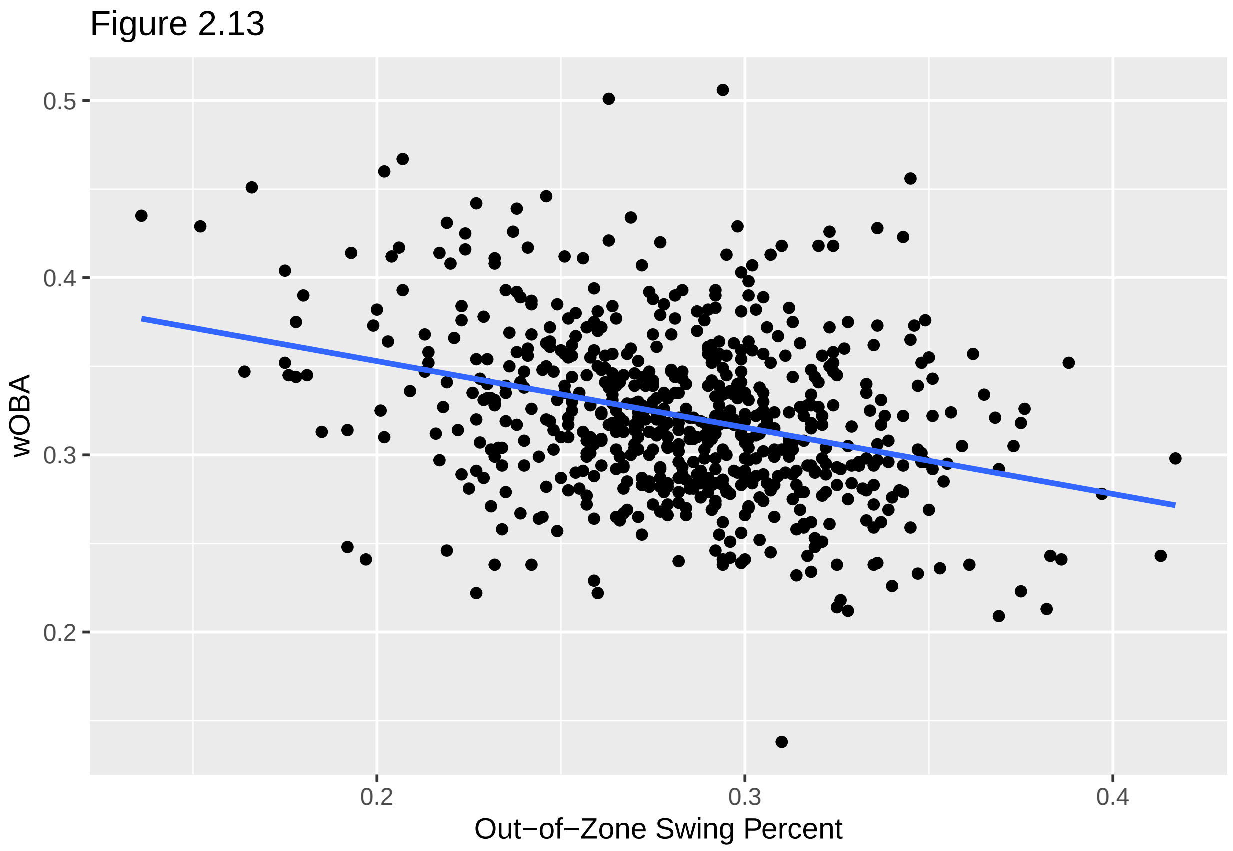
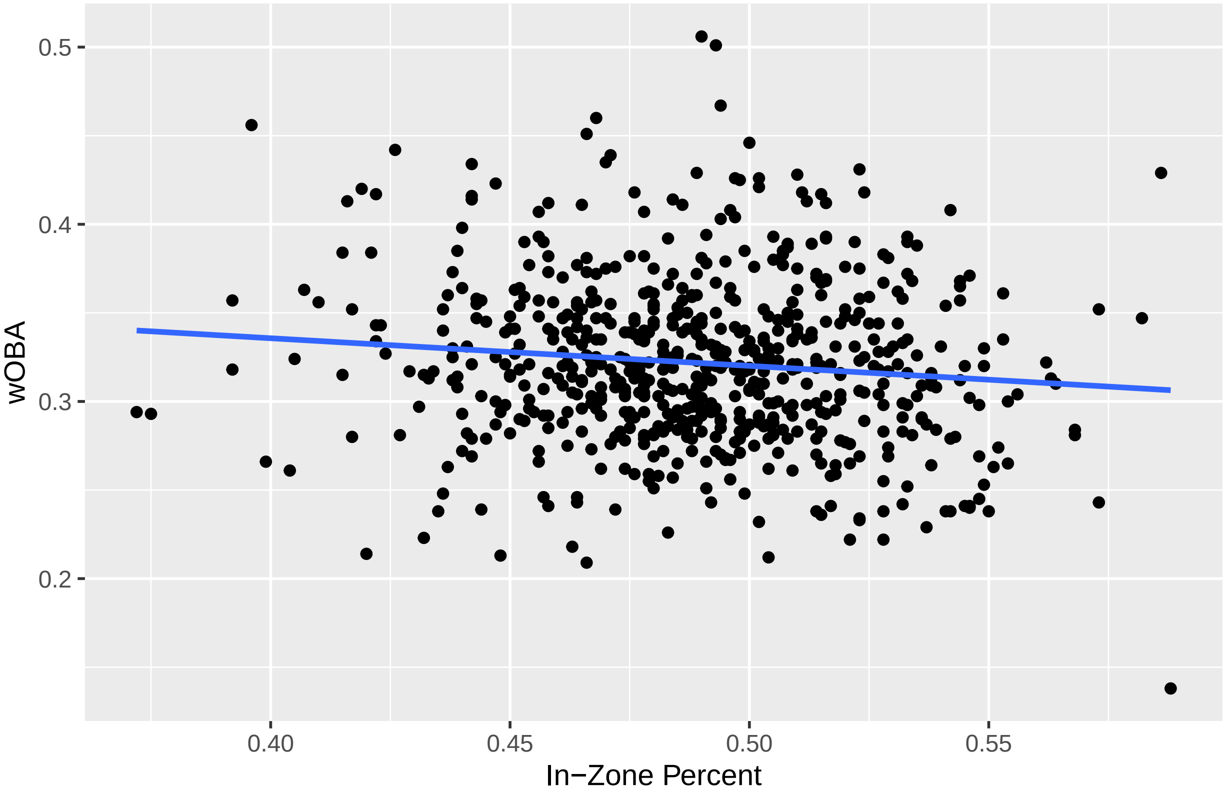


Figure 2.13 displays a negative linear relationship between wOBA and out-of-zone swing percent. This is as expected, as more out-of-zone swings from a pitcher is associated with better performance.

In-Zone Percent:

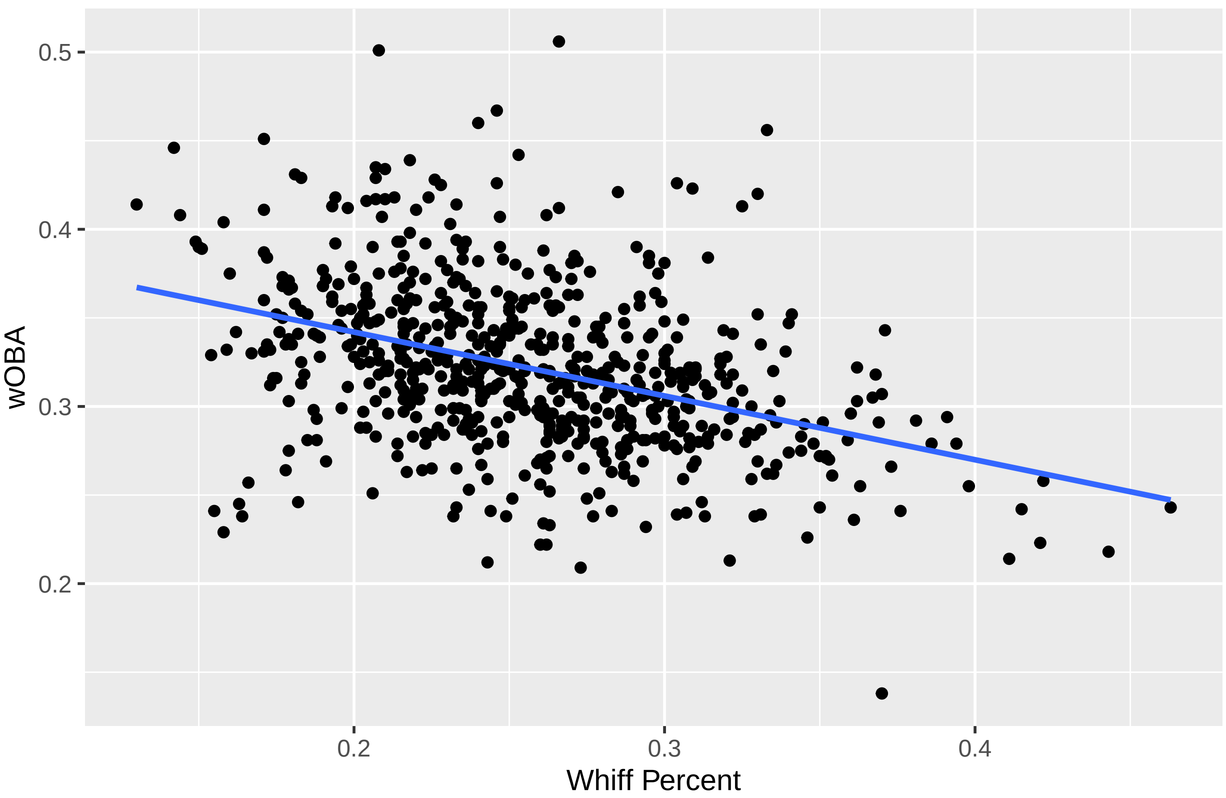
Figure 2.14



According to Figure 2.14, in-zone percent appears to have a very slight linear relationship with wOBA. It is generally assumed that throwing more pitches in the strike zone yields better results, but that also indicates that a pitcher may be giving hitters easier pitches to hit. Thus, this relationship being extremely weak makes sense.

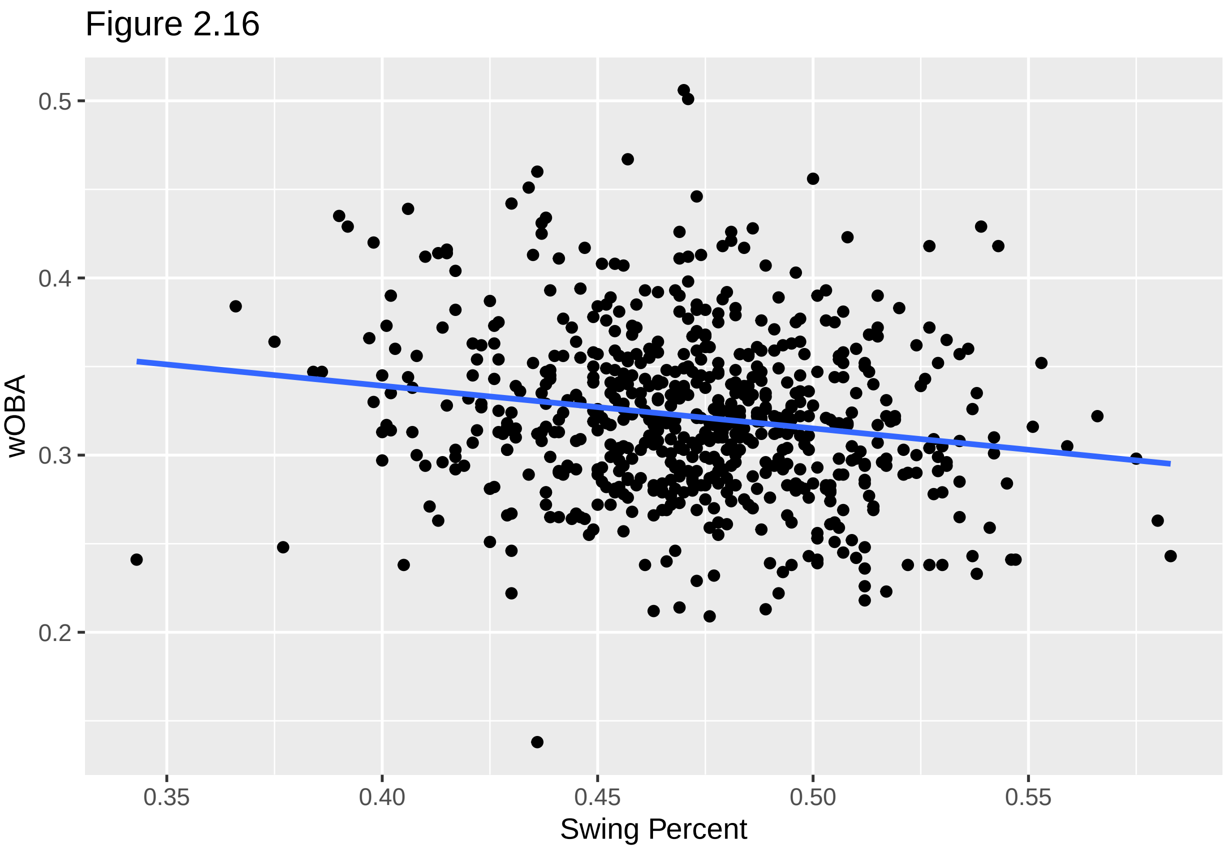
Whiff Percent:

Figure 2.15



There appears to be a negative linear relationship between whiff percent and wOBA, per Figure 2.15. This is expected, as a higher whiff rate is generally associated with better pitcher performance.

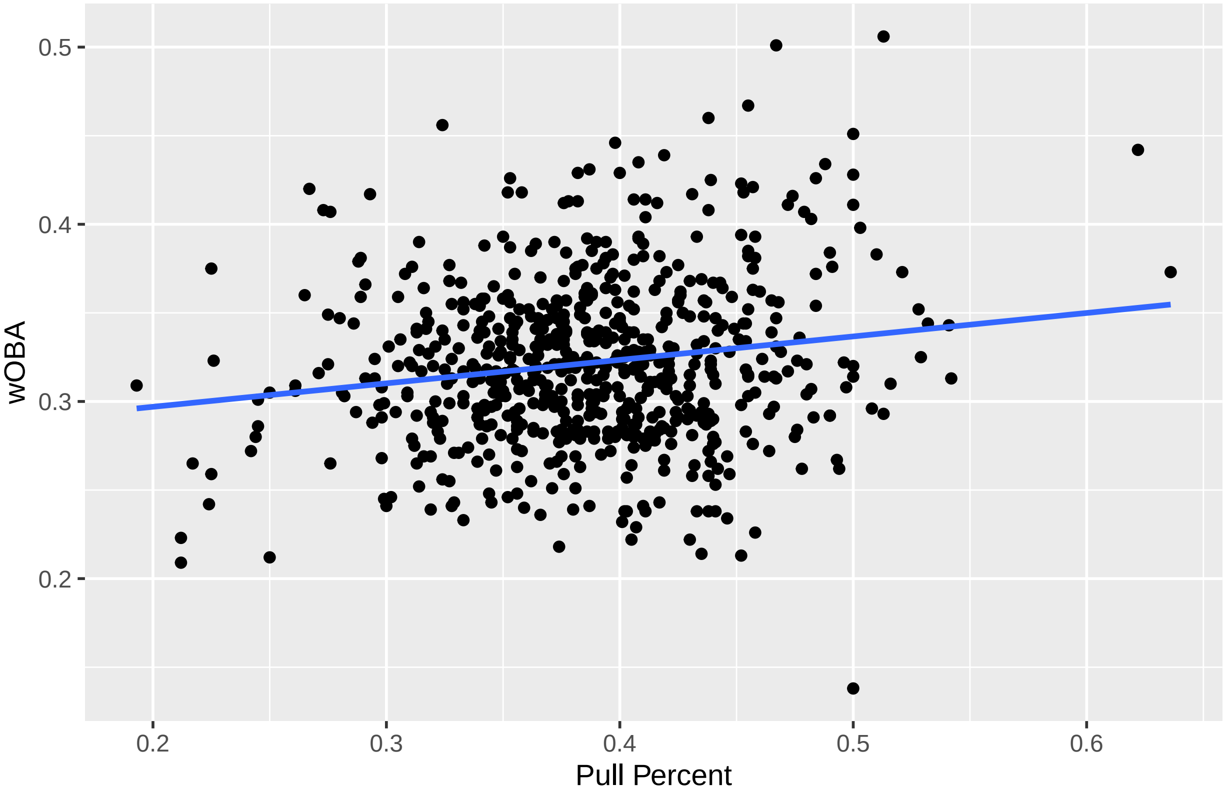
Swing Percent:



There appears to be a slight negative linear relationship between swing percent and wOBA according to Figure 2.16. Swing percent is similar to in-zone percent in that it's not as strongly associated with pitcher success, however, this does seem plausible.

Pull Percent:

Figure 2.17



For our last variable to check, Figure 2.17 shows a positive linear relationship between wOBA and pull percent. This makes sense, as pull percent is similar to hard hit percent and barrel percent in that it is generally associated with worse pitcher performance.

Section 3: Method 1 (KNN)

Section 3.1 - KNN Introduction

Now that our data has been cleaned and exploratory data analysis was performed, we can perform our first method, K-Nearest Neighbors (KNN). We are starting with this method, as it simply uses our provided data set to create predictions. This should provide us with serviceable predictions that we can compare to the accuracy our future, more comprehensive methods' predictions.

Section 3.2 - KNN Method

KNN is a predictive algorithm that uses our given data set to make predictions. These predictions are made using validation data created by 10-fold CV. KNN finds the most similar rows to a specific row in the data set, chooses the K closest features for the row, and averages their wOBA values to predict wOBA. This is done for every row in the data set. In order to find our K, we performed a method called tuning, which involved running KNN for many different possible K values and selecting the K that gives us the best predictive accuracy. After performing this, we ended up with K equal to 6.

The features' similarity is determined by a distance metric. We chose Gower's Distance because its formula is simple to understand and works for all types of features that have different ranges. This is especially helpful in our cases since we are using variables with quite different ranges. This formula computes the distance by comparing the values of each feature in our row to every other row in the set, also incorporating the range of the feature. For each row in our data set, the 6 other features with the lowest Gower's Distance are deemed the most similar, and their average wOBA value is our prediction for the row.

Section 3.3 - KNN Results

Table 2: Figure 3.3.1: RMSE for Optimal K of KNN

	K	RMSE
6	6	0.0284902

The predictive results of KNN are shown in Figure 3.3.1 above. We see a test RMSE of 0.028. This means that the average distance between the predicted wOBA from KNN and the actual wOBA values is 0.028. Considering our mean wOBA from the data set is 0.322, that is not a bad RMSE. A change in RMSE by roughly 0.03 is not the most significant, however we want this RMSE to be as low as possible in order to most accurately predict wOBA. With KNN, a rare few pitchers could have predicted wOBA values that are as far off as 0.05 or more, and this is where it approaches the territory of our modeling an average pitcher as a good pitcher or the other way around. This predictive accuracy is not as high as desired.

Additionally, it is difficult to interpret these results outside of predictive accuracy. It is more difficult to tell what statistics specifically have an impact on wOBA. Therefore, it would be best to explore other methods.

Section 4: Method 2 (Elastic Net)

Section 4.1 - Elastic Net Introduction

Next, we move on to Elastic Net. This is a complex method that should generate better predictive accuracy than KNN. This is particularly beneficial because we have correlated features, as shown in Figure 4.1.1 below. If features are extremely correlated, this method would be necessary, as general LSLR would blow up our coefficient estimates.

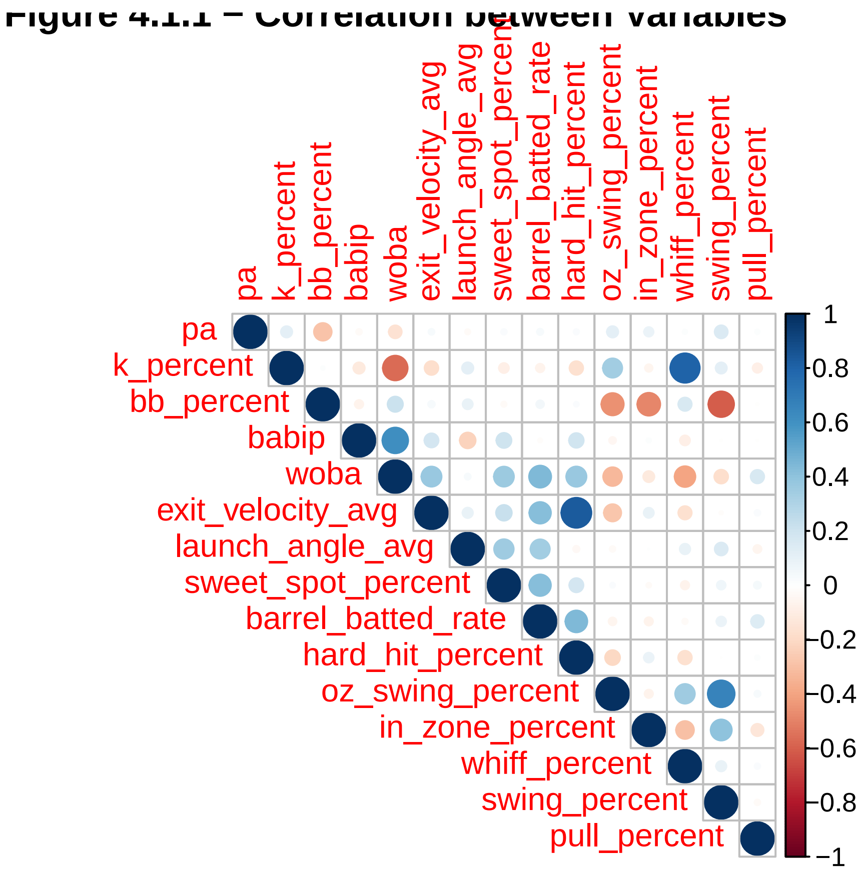


Figure 4.1.1 – Correlation between variables

All boxes that contain large circles (not including the left-most box of each row, since that is checking a feature with itself) indicate correlation between two features. While there are not many cases and nothing that would ruin LSLR, there are a few cases of high correlation between variables that warrant the use of elastic net, with an example being average exit velocity and hard hit percentage.

Section 4.2 - Elastic Net Method

Elastic Net is a more complex way of running a least squares linear regression model. In linear regression, we want to use estimates that minimize the residual sum of squares (RSS). For elastic net, we want to use estimates that minimize the sum of the RSS and a penalty term. This penalty term attempts to balance bias and variance, as we would like as low bias and variance as achievable.

This process involves two terms: α and λ. The α term determines how much “selection” to use, which is how willing we are to set coefficients equal to zero, thus eliminating terms from the model. This operates on a scale of 0 to 1, where zero does not allow for any selection (also called ridge regression), and one fully allows selection (also called lasso regression). The λ term determines how large the penalty term that was introduced earlier should be. In other words, it controls how much “shrinkage” of the coefficients we allow. In our case, since our response variable and by proxy RSS is quite small by nature, it is best to use a λ that is also quite small.

With that in mind, we create a model that minimizes the sum of RSS and penalty term by simply testing out all plausible combinations of α and λ We create a model for each plausible combination and choose the one that yields the lowest RMSE.

Section 4.3 - Elastic Net Results

Table 3: Figure 4.3.1: α, λ, and RMSE of Elastic Net Model

	α	λ	RMSE
98	0.97	3e-04	0.0174707

Figure 4.3.1 outputs the important results from this method. An initial observation is that our λ is equal to 0.0003, which indicates that some features may have been removed (coefficient equal to 0) due to the model deeming them insignificant.

The test RMSE of 0.017 is what will give an idea of whether elastic net is more beneficial in terms of accomplishing our goal of prediction. Using elastic net, the average distance between the predicted wOBA and the actual wOBA is 0.017. This value is significantly (61%) lower than our RMSE for KNN, meaning it has much higher predictive accuracy than KNN. This alone implies that elastic net is a more preferable method than KNN, but it should also be noted that we are also given coefficients for each feature with this method. Thus, we now have an idea of how our individual features impact wOBA. These coefficients are displayed in Figure 4.3.2.

Table 4: Figure 4.3.2 - Coefficients of Elastic Net Model

Feature	Coefficient
(Intercept)	0.1425601
pa	-0.0000065
k_percent	-0.3796542
bb_percent	0.2847915
babip	0.5659709
exit_velocity_avg	0.0000000
launch_angle_avg	0.0005600
sweet_spot_percent	0.0254179
barrel_batted_rate	0.5974052
hard_hit_percent	0.0100174
oz_swing_percent	0.0000000
in_zone_percent	0.0000000
whiff_percent	0.0000000
swing_percent	-0.0373392
pull_percent	0.0582142

As warned out above, some features have coefficients equal to 0, meaning they are not incorporated into our model. Those features are average exit velocity, out-of-zone swing percent, in-zone percent, and whiff percent. This tells us that those features are likely not needed to predict wOBA if we have all of the other statistics in the data set. They are not statistics to look for when scouting pitchers to approach.

The only issue with elastic net is that these coefficients are quite small, which is to be expected given our wOBA has a small range. This makes our coefficients harder to interpret. We can estimate how much an increase in a feature affects wOBA, but these will not be easy to understand. This is especially an issue because the decisions we are making involve coaches, players, and front office members that do not know much beyond the statistics themselves.

Ultimately, this method yields a much better RMSE and thus predictive accuracy, and we are now able to see how each feature impacts wOBA according to the model. This method would be a great choice. It is worth exploring one final method, however, as there is a specific method that would allow us to explore the effects of these individual features much more and have an easier time actually determining which pitchers we want to go after. This would make it easier to convince coaches and staff that these players are worth signing, as we can point out the most important statistics to look out for (have a small wOBA).

Section 5: Method 3 (Random Forest)

Section 5.1 - Random Forest Introduction

The final method to be used is random forest. This method allows us to use feature importance to examine how features are impacting the model.

Section 5.2 - Random Forest Method

To understand a forest model, you must first understand regression trees. These are models that start with all features in the data set and branch out using splits. At each split, the data set is split into two depending on whether they satisfy a condition related to a feature. For numeric features, it is usually a threshold of the value of the feature. Trees contain nodes that contain different amounts of observations from the data set. Trees start with a root node that contains the entire data set, and then it splits into two "leafs". These leafs contain the subset of the observations that satisfy the split, and then the next set of splits will be performed solely on these leaves and their respective observations. An example of a split is "exit_velocity_avg < 0.95", where one node contains the sub-population of our data that satisfies this condition, and the other contains the rest of the population that does not satisfy this condition.

A problem with trees is that they over-fit the data set since they are trained on our data, and thus we should use a forest model. Forests consist of many trees that are all grown are different versions of the training data. This data is contained by bootstrap sampling, which is sampling our data set with replacement so that not every observation is contained. This results in different samples that will be used to train the trees.

There are two types of forests: bagged forests and random forests. We are using a random forest. The difference between them is that bagged forests consider all features at each split for each tree across the forests (although categorical variables can only be split on once), while our random forest considers only three (the rounded square root of the number of features, rounded down if needed). This will make our process of growing the forest quicker and encourage feature diversity among the trees, which will allow us to properly consider all features in the model.

To expand on the benefit of this model, we can examine the feature importance once all these steps to create the model are performed. The simple way to explain importance is that all of these bootstrap samples that the trees in the forest are trained on exclude some observations. We call these out-of-bag (OOB) observations. These are used as test data to evaluate each tree. This allows us to find prediction metrics like RMSE/MSE and determine the importance of each feature, which will be shown for our model below.

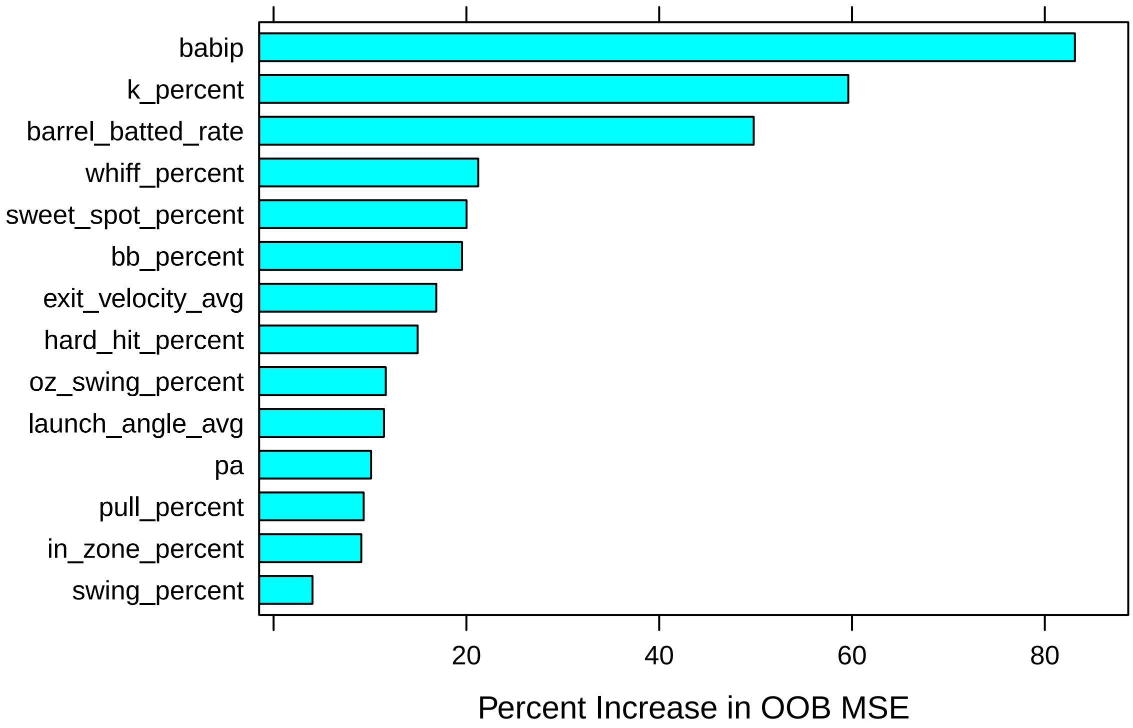
Section 5.3 - Random Forest Results

Table 5: Figure 5.3.1: RMSE of Random Forest

x
0.0247303

Figure 5.3.1 displays the test RMSE of our random forest, which is 0.025. For our forest model, the average distance between the predicted wOBA and the actual wOBA is 0.025. This is better than our KNN result, but worse than elastic net. While this predictive accuracy is not bad, the primary value of the random forest model comes from feature importance. The feature importance can be seen in Figure 5.3.2 below.

Figure 5.3.2: Importance



This figure shows us which statistics have the highest importance, or cause a large increase in MSE when added to the model. We can see that BABIP, strikeout percent, and barrel rate are quite significantly the three features with the most importance. This will provide us a lot of value when looking for pitchers. Especially when looking at players in college, high school, the minor leagues, or in leagues in other countries, it will be very beneficial to examine these statistics as influential players in wOBA prediction. Often our data set given all of these potential players will be missing some features or too large to operate some of these models, and it would therefore be beneficial to look at the features with the largest importance shown above.

Thus, our random forest model plays an important role in our goal of acquiring the best possible pitchers we can get in a cost-effective way. Despite its predictive accuracy being lower than elastic net, this model can play a supplemental role in finding pitchers to analyze.

Conclusion

In conclusion, the KNN method is not something that should be used to predict wOBA going forward due to lower predictive accuracy than other methods and no upside in terms of interpretation. However, both elastic net and random forests should be used by the organization when looking for pitchers, each for their own reason. Elastic net clearly provided us with the highest predictive accuracy of the three methods, and thus that should be used to predict wOBA when evaluating pitchers. Random forests, on the other hand, provide an easy to understand output that shows what features in particular play a role in wOBA prediction. This is very helpful when dealing with coaches and players who want to understand our process better, and it is also helpful when our data is limited while scouting players abroad or in lower levels where we may not be provided all the data we need to predict wOBA.

Therefore, my recommendation would be to first divide your prospective pitchers based on what data we have for them. For players with data on all of the features we have in this data set (likely MLB players or

Division 1 college players), we can directly use our elastic net model to predict wOBA for their next season in a similar league. However, for pitchers with limited data, we should look out for promising values in BABIP, strikeout percent, and barrel rate as indicators of success or talent that we can develop. Batted ball data (data involving exit velocity or launch angle) is a common set of data that not all leagues provide since it requires the use of expensive technology, and thus looking at BABIP or strikeout percent will be extremely helpful when scouting players. This philosophy should improve the success rate of scouting pitchers in the coming years, and it will also allow us to make more intelligent free agent transactions or trades.

Glossary

Plate Appearances - *“pa”* - The total amount of batters faced by the pitcher in 2023.

Strikeout Percent - *“k_percent”* - The percentage of batters faced who the pitcher struck out in 2023.

Walk Percent - *“bb_percent”* - The percentage of batters faced who the pitcher walked in 2023.

BABIP (Batting Average on Balls in Play) - *“babip”* - The batting average on all balls put in play against the pitcher in 2023. That is, the number of batted balls against the pitcher resulting in hits over the number of all batted balls put in play against the pitcher.

wOBA (Weighted On-Base Average) - *“woba”* - The wOBA of all opposing hitters on plate appearances against the pitcher in 2023. The wOBA comes from a formula that is shown here. For pitchers, this measures roughly how well hitters perform against each pitcher, with lower wOBA indicating better pitcher performance.

Average Exit Velocity - *“exit_velocity_avg”* - The average exit velocity, in miles per hour, of balls put in play against the pitcher in 2023. In other words, on average, how fast was the baseball travelling upon being struck by the opposing hitter’s bat.

Average Launch Angle - *“launch_angle_avg”* - The average launch angle, in degrees, of balls put in play against the pitcher in 2023. In other words, at what angle did the ball exit from the opposing hitter’s bat for each ball put in play.

Sweet Spot Percent - *“sweet_spot_percent”* - The percentage of batted balls hit between 8 and 32 degrees of launch angle against the pitcher in 2023.

Barrel Rate - *“barrel_batted_rate”* - The percentage of balls put in play against the pitcher in 2023 that are considered “barrels”. A barrel is defined by MLB here, but in simple terms, a batted ball is considered a “barrel” if it has an exit velocity and launch angle that result in a high likelihood of extra-base hits (double, triple, home run) among previously recorded batted balls with similar data.

Hard Hit Percent - *“hard_hit_percent”* - The percentage of balls put in play against the pitcher in 2023 that are hit with an exit velocity of 95 mph or greater.

Out-of-Zone Swing Percent - *“oz_swing_percent”* - The percentage of balls thrown outside of the strike zone that the hitter swung at for the pitcher in 2023.

In-Zone Percent - *“in_zone_percent”* - The percentage of all pitches thrown by the pitcher in 2023 that were located in the strike zone.

Whiff Percent - *“whiff_percent”* - The number of times a batter swung and missed against the pitcher in 2023 divided by the number of times a batter swung at all against the pitcher in 2023.

Swing Percent - *“swing_percent”* - The percent of pitches thrown by the pitcher in 2023 that garnered a swing by the opposing batter.

Pull Percent - *“pull_percent”* - The percent of balls in play against a pitcher in 2023 in which a batter hits the ball to the side of the field from which they bat. In other words, a pull is considered when a right-handed hitter hits it to the left side, and a left-handed hitter hits it to the right side.

Works Cited

pitchers. Retrieved April 8, 2024 from baseballsavant.mlb.com.