Measuring and Modeling Batted Ball Quality

WSU Mathematics and Statistics department

Introduction

Sports analytics has become very popular in recent years. Sports teams are looking for ways to gain a competitive advantage using data. There are also plenty of online news outlets that use data to tell stories about sports. Many of these teams and news outlets use Statcast data. An example would be from the Hardball Times where the research involved predicting hit outcomes using the Statcast data (Petti). The study that we conducted uses the same data but explores a different question.

The question that we explored only considered pitching variables in predicting the quality of a hit ball. This idea led us to explore the following questions. What is well hit ball? Can we explain a pitch in a different way? What pitching characteristics lead to a well hit ball? The answers to these questions can influence the strategy behind how a pitcher attacks a hitter. For example, if a team knows what type of pitch makes one hit the ball poorly, then a team should throw those types of pitches more often. Another example would be the hitting team learning what types of pitches certain players hit very well. This may lead to instructing the hitters for what types of pitches to look swing at.

Methods and materials

Statcast technology has given a rich source of data for every pitch in the Major League Baseball Association (MLB). This data is available at *baseballsavant.mlb.com*. This URL is where anyone can find the data that we used. The data has an observation for every single pitch. There are a total of 60 variables for each player. Variables consisted of hit speed, hit distance, the amount of break on a pitch, the type of pitch, the exact coordinates of where the pitch crossed the plate, and many more. Once the data was obtained, we made a subset of the data with only hits that were in the field of play. We had a separate data set for each of the top ten players, according to the Wins Above Replacement (WAR), statistic: Brian Dozier, Jose Altuve, Josh Donaldson, Kris Bryant, Kyle Seager, Manny Machado, Mike Trout, Mookie Betts, Nolan Arenado, and Robinson Cano (baseball-reference). These were the final data sets used for all of the statistical analysis.

The statistical analysis of this data can be broken up into three main parts: understanding what a well-hit ball is, understanding what a pitch is, and modeling a hit ball using pitching variables. The main idea was to first understand what constitutes as a hit baseball, and then to understand what types of pitches impacted batted ball quality. The rest of this section will discuss the methods used to investigate these three main parts of the statistical analysis.

Understanding what a well hit-ball and a pitch was investigated using principal components analysis. Both principal components analyses scaled each variable to have mean zero. This ensures that we are not computing components based off of an arbitrary choice of scaling

(James). The principal components analysis for hitting used three variables: hit speed, hit distance, and hit angle. The principal components analysis for the pitching variables consisted of 24 pitching variables: start speed, x0, z0, spin direction, spin rate, break angle, break length, pfx_x, pfx_z, px, pz, hc_x, hc_y, vx0, vz0, ax, ay, az, sz_top, sz_bot, effective speed, release spin rate, and release extension. We used the loadings, scores, and the proportion of variance explained to understand the results.

For modeling a hit ball using pitching variables, we fit five models: full linear model, forward selection, backward selection, ridge regression and the lasso. These models are a blend of standard linear models with subset selection, and linear models with shrinkage methods. We used 26 pitching variables to predict hit speed. The full linear model fit is shown below:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{26} X_{26} + \epsilon$$

Forward selection is subset selection method that starts with no predictors in a model. At each iteration, it adds the variable that gives that model the most additional improvement. The algorithm will iterate through all of the variables. Once complete, the algorithm will return the model with the lowest AIC. Backward selection is very similar to forward selection, however, it begins with the full least squares model containing all predictors. The AIC is given by

$$AIC = \frac{1}{n\sigma^2}(RSS + 2d\hat{\sigma}^2),$$

where, for simplicity, we have omitted an additive constant (James).

Ridge regression was also fit on all ten players using all 26 variables to predict hit speed. Ridge regression is very similar to a least squares model, except that it has an additional tuning parameter and shrinkage penalty. "Ridge regression is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity. In particular, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2,$$

Where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately" (James). We selected a value of λ by using 10-fold cross validation on the training set.

The final model was the lasso. The lasso model is similar to ridge regression, except that it allows some coefficients to take on the value of 0. Whereas ridge regression will include all 26 predictors in the model. This feature selection will allow us to determine what pitching characteristics explain the quality of a hit ball. "The *lasso* is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

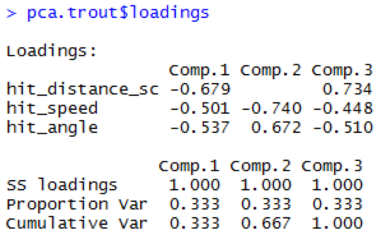
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| = RSS + \lambda\sum_{j=1}^{p}\left|\beta_j\right|.\text{” (James).}$$

We assessed model performance by comparing the R^2 values on the 35% training data. All of the models were built using the 65% training data. This is a form of cross-validation where we use a training set to build models and a test set to compare model performance.

Results

This results section will quickly summarize the output from the three main statistical analyses: PCA on pitching variables, PCA on hitting variables, and the modeling performance on hit speed. Each analysis will use data visualizations to summarize results throughout. The principal components analysis results were only done on Mike Trout. Whereas the modeling results will be discussed using the top ten players according to the WAR statistic for position players from baseball-reference.com.

The PCA on the 3 hitting variables of hit speed, hit distance, and hit angle had interesting results. The R code output below in Figure 1 gives us the loadings and cumualitve variance explained:



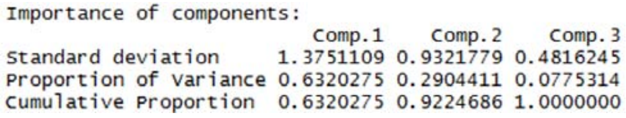
```
> pca.trout$loadings

Loadings:
                Comp.1 Comp.2 Comp.3
hit_distance_sc -0.679         0.734
hit_speed       -0.501 -0.740 -0.448
hit_angle       -0.537  0.672 -0.510

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
```

Figure 1: Principal component analysis on hitting variables – output

This output shows us that all three variables were pretty equal on the first component when explaining the variance between the three variables. In the second loading, we can observe hit speed and hit angle are the two variables that explain most of the remaining variance. The cumulative variance is 0.922 after the first two components, as shown below in figure 2. This means that only 92% of the variance can be summarized using the first two principal components. Figure 3 below is a data visualization that plots the first two components and is colored by hit outcome or event.



```
Importance of components:
                          Comp.1    Comp.2    Comp.3
Standard deviation     1.3751109 0.9321779 0.4816245
Proportion of Variance 0.6320275 0.2904411 0.0775314
Cumulative Proportion  0.6320275 0.9224686 1.0000000
```

Figure 2: Importance of components for PCA hit

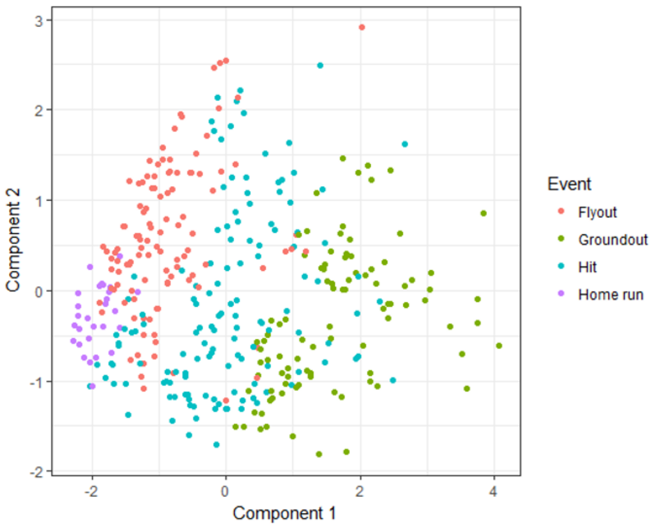
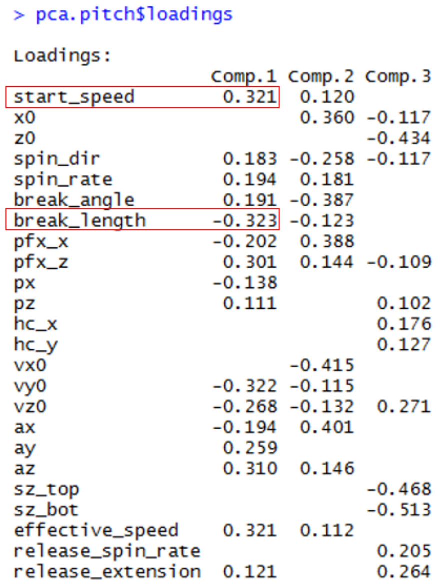


Figure 3: First two scores plotted by event color

This plot reveals very interesting results. We can observe clusters within the event type naturally occurring within the first two components. For example, home runs mostly cluster together in the lower left-hand corner of the plot, as seen in purple. We can also observe strips of clusters forming for fly outs, groundouts, and hits. If one were to model home runs, then using the first component as a response variable may be of interest. With that being said, we decided that with a cumulative variance of 0.92, we should just select a single response variable. We chose to use hit speed for the response variable.

The next section is the PCA on the 24 pitching variables. The R-code below in figure 4 shows the first loadings for the first three components:



```
> pca.pitch$loadings

Loadings:
                  Comp.1 Comp.2 Comp.3
start_speed        0.321  0.120       
x0                        0.360 -0.117
z0                              -0.434
spin_dir           0.183 -0.258 -0.117
spin_rate          0.194  0.181       
break_angle        0.191 -0.387       
break_length      -0.323 -0.123       
pfx_x             -0.202  0.388       
pfx_z              0.301  0.144 -0.109
px                -0.138              
pz                 0.111         0.102
hc_x                             0.176
hc_y                             0.127
vx0                      -0.415       
vy0               -0.322 -0.115       
vz0               -0.268 -0.132  0.271
ax                -0.194  0.401       
ay                 0.259              
az                 0.310  0.146       
sz_top                          -0.468
sz_bot                          -0.513
effective_speed    0.321  0.112       
release_spin_rate                0.205
release_extension  0.121         0.264
```

Figure 4: The loadings for the first three components of the PCA pitching

The two variables and loadings outlined in red are start_speed and break_length. In other words, the first component gives most of the weight to these two variables. This is interesting, because the first component seems to be corresponding to whether or not the pitch is a fastball type or off-speed type pitch. For example, a curveball is usually slower and has more break than a four-seam fastball. To investigate whether or not this component was splitting the data based on whether or not the pitch was a fastball type or off-speed type, we plotted the first two components and colored the data points by whether or not it was a fastball type or off-speed type.

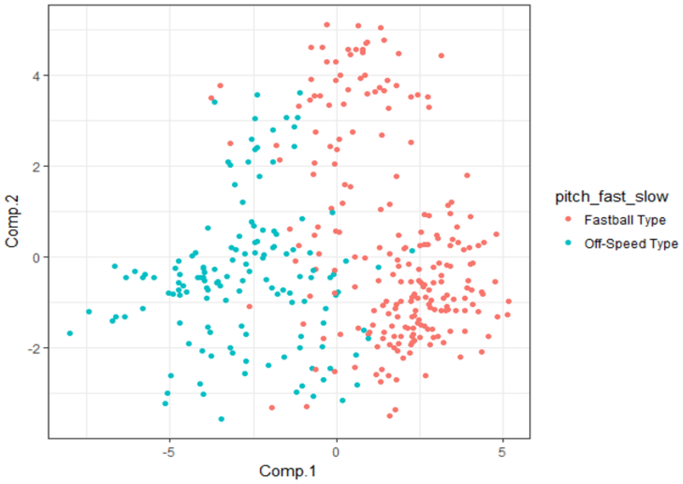
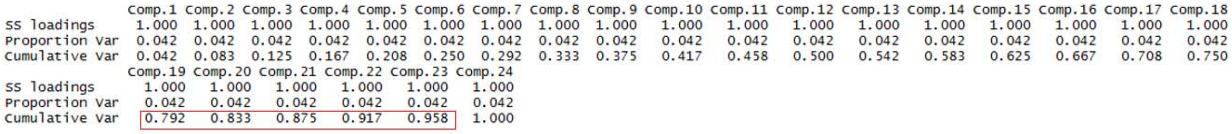


Figure 5: The first two components plotted with color on whether or not the pitch was a fastball type or off-speed type

Figure 5 reaffirms the results from the loading vector in Figure 4. This tells us that the break length and the speed of the pitch are related to one another. Specifically, a pitch with less break will go faster than a pitch with more break.

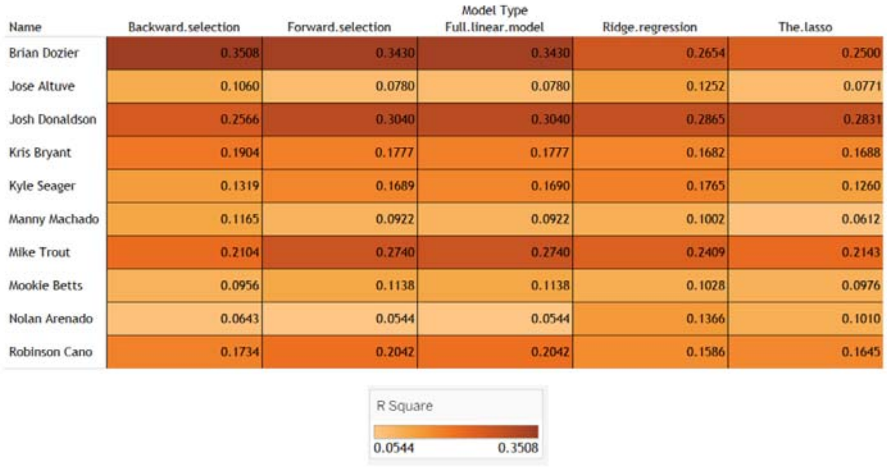
One of the main ideas behind a PCA, is to reduce the number of variables in a data set. This would have been very applicable for our modeling scenario of using 26 variables, however, the cumulative proportion explained after the components was not high enough to do so. Figure 6 shows that components and their respective cumulative proportions.



```
               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17 Comp.18
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000   1.000   1.000   1.000   1.000   1.000   1.000   1.000   1.000   1.000
Proportion Var  0.042  0.042  0.042  0.042  0.042  0.042  0.042  0.042  0.042   0.042   0.042   0.042   0.042   0.042   0.042   0.042   0.042   0.042
Cumulative Var  0.042  0.083  0.125  0.167  0.208  0.250  0.292  0.333  0.375   0.417   0.458   0.500   0.542   0.583   0.625   0.667   0.708   0.750
               Comp.19 Comp.20 Comp.21 Comp.22 Comp.23 Comp.24
SS loadings      1.000   1.000   1.000   1.000   1.000   1.000
Proportion Var   0.042   0.042   0.042   0.042   0.042   0.042
Cumulative Var   0.792   0.833   0.875   0.917   0.958   1.000
```

Figure 6: The cumulative proportions of the components for PCA pitching

The final analysis that was conducted was the five modeling techniques for the top ten players according to the WAR statistic for positon players from baseball-reference.com. Figure 7 below shows us the modeling results for each model and each player. Each box is colored by the cross-validated R^2 value.



Name			Model Type		
	Backward.selection	Forward.selection	Full.linear.model	Ridge.regression	The.lasso
Brian Dozier	0.3508	0.3430	0.3430	0.2654	0.2500
Jose Altuve	0.1060	0.0780	0.0780	0.1252	0.0771
Josh Donaldson	0.2566	0.3040	0.3040	0.2865	0.2831
Kris Bryant	0.1904	0.1777	0.1777	0.1682	0.1688
Kyle Seager	0.1319	0.1689	0.1690	0.1765	0.1260
Manny Machado	0.1165	0.0922	0.0922	0.1002	0.0612
Mike Trout	0.2104	0.2740	0.2740	0.2409	0.2143
Mookie Betts	0.0956	0.1138	0.1138	0.1028	0.0976
Nolan Arenado	0.0643	0.0544	0.0544	0.1366	0.1010
Robinson Cano	0.1734	0.2042	0.2042	0.1586	0.1645

Figure 7: The modeling results

We can observe that the cross-validated R^2 values ranged between 0.0544 and 0.3508. When we compare models across players, there doesn't seem to be a stand-alone best model. One observation we can make is that the accuracy of the models seems to be somewhat consistent within a particular player.

The last part of the statistical analysis was to view the lasso coefficients for each player. The lasso coefficients naturally perform feature selection, and will tell us which variables lead to predict hit speed. Figure 8 below shows the lasso coefficients in a bar plot for Mike Trout.

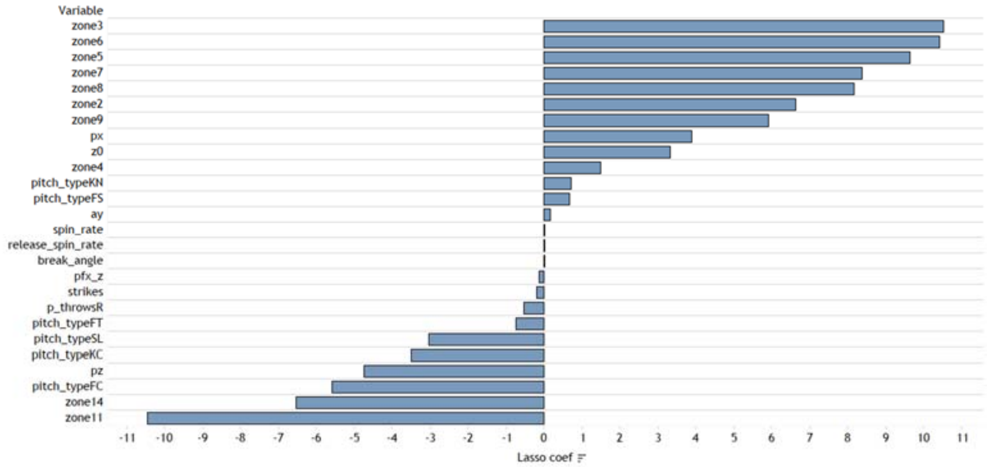


Figure 8: Lasso coefficients for Mike Trout

By observing Figure 8, we can see that the dummy variables zone 3, zone 6, and zone 5 lead to a higher predicted hit speed. This means that if a pitch was in zone 3, zone, 6, or zone 5 the model would predict a higher hit speed. Figure 9 below shows the zone locations.

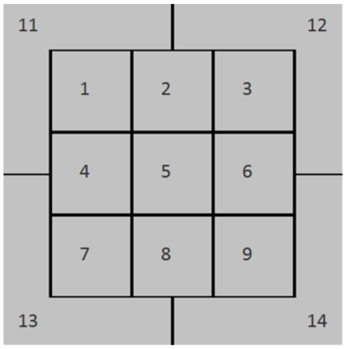


Figure 9: Zone locations

The variables that lead to the lasso model predicting a lower hit speed for Mike Trout are zone 11, and zone 14. This means that pitches within the uppermost left zone and bottommost right zone were predicted to give Mike Trout a lower hit speed. Figure 10 below is a data visualization of every pitch hit in play colored by hit speed for Mike Trout.

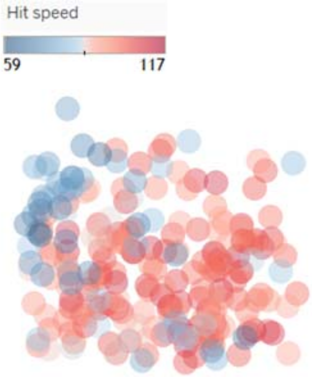


Figure 10: Pitch location colored by hit speed for Mike Trout

Figure 10 goes along with the lasso coefficient plot from Figure 8 in that the three variables zone 3, zone 6, and zone 5 lead to the highest predicted hit speed. By looking at Figure 10 and Figure 9, we can see that there seems so be a darker red color emerging within zones 3, 6, and 5.