Predicting Batted Ball Outcomes in Major League Baseball

Abstract

Statcast is a radar tracking technology implemented in 2015 in MLB Ballparks with the goal of measuring various metrics in baseball games. The data is publicly available and has become key in assessing player performance. Exit velocity and launch angle are key metrics that are used in assessing batter performance. Exit velocity and launch angle for every batted ball in early April was downloaded from the Statcast website. A quasibinomial model was fit using exit velocity, launch angle and squared launch angle. Exit velocity, squared launch angle, and the interaction between exit velocity and squared launch angle were found to significantly predict an outcome of a hit or an out. This information may be used to develop players and evaluate performance.

Introduction

Major League Baseball introduced its Statcast tracking technology to all 30 MLB ballparks before the 2015 season, allowing not only teams but fans access to vast amounts of data about America's pastime. Statcast tracks almost every metric a fan or front office member could want: from pitch spin rate and release point to baserunner and fielder run speed to two of the most cited statistics in launch angle and exit velocity. Using these quantities, player evaluation and development can be greatly improved compared to previous methods.

Hitting is one of the most crucial elements for any player evaluation, from a fan choosing a player for their fantasy roster or a general manager trying to correctly identify who is deserving of a multimillion-dollar contract. Previous methods such as batting average can't quantify how lucky or unlucky a player has been, but by using launch angle and exit velocity, probabilities of each batted ball can be calculated and allowing for some calculation of luck. Launch angle and exit velocity have not only helped in player evaluation, but also with player development. Several prominent players such as Josh Donaldson, Justin Turner, and J.D. Martinez have attributed their performance in the last few years to increasing the launch angle on their batted balls. In addition, average launch angle has increased from 10.1 in 2015 to 11.0 in 2017 and shows no signs of decreasing anytime soon.

Methods

The data was obtained from the Statcast website, with every batted ball from the beginning of the season through April 6 included in the sample. Prior to analysis, variables for centered exit velocity, squared launch angle, z-score launch angle, and squared z-score launch angle were created. An indicator for a successful hit was created, which was the response variable used. In addition, all bunts were removed since they are fundamentally different than a full swing, bringing the total number of batted balls from 5843 to 5803. Exploratory analysis was performed examining the probability of a base hit and exit velocity and launch angle separately, (Appendix models 1&2).

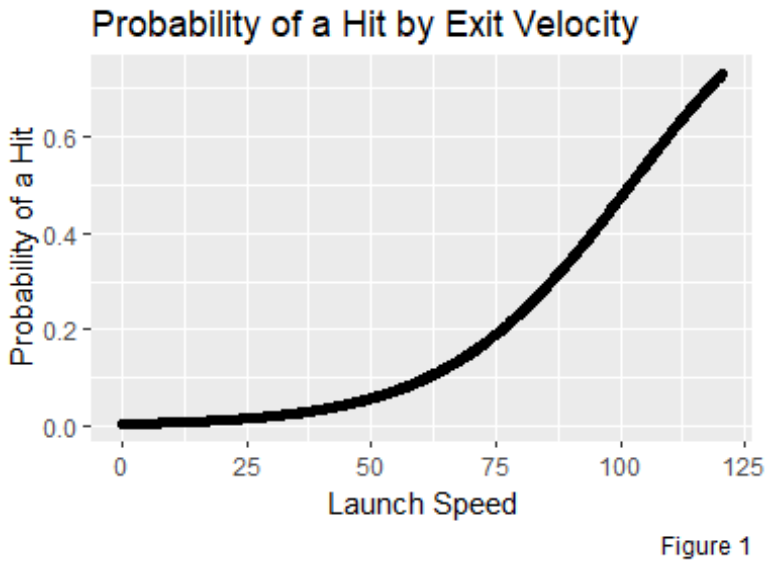


Figure 1

Figure 1 shows the relationship between exit velocity and batted ball outcome. Increased exit velocity appears to increase the probability of a base hit.

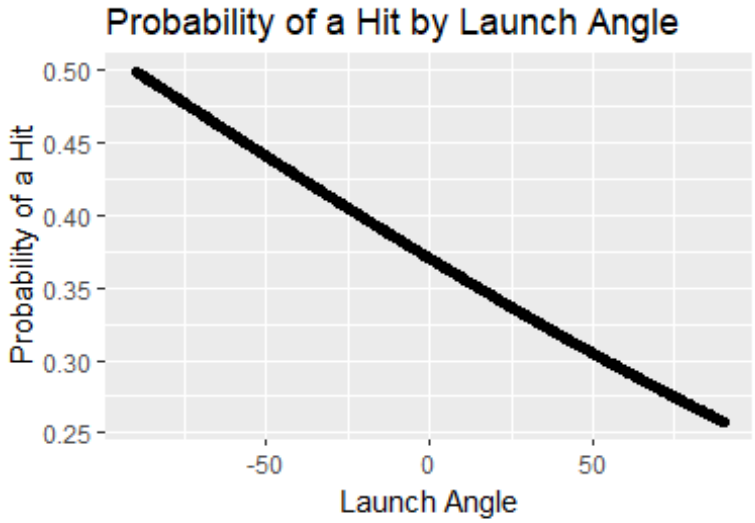


Figure 2

The probability of a hit appears to decrease with increased launch angle, as seen in figure 2. However, the relationship appears to have a potential quadratic effect (Appendix model 3), as seen in figure 3, which is included in the final model.

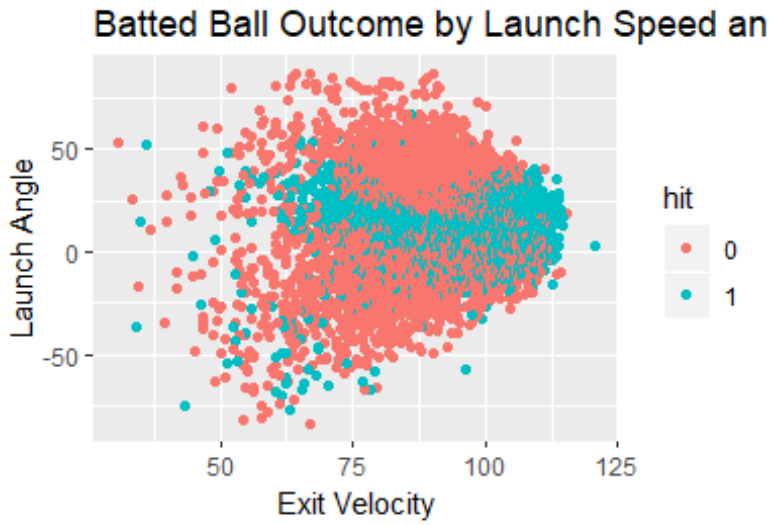


Figure 3

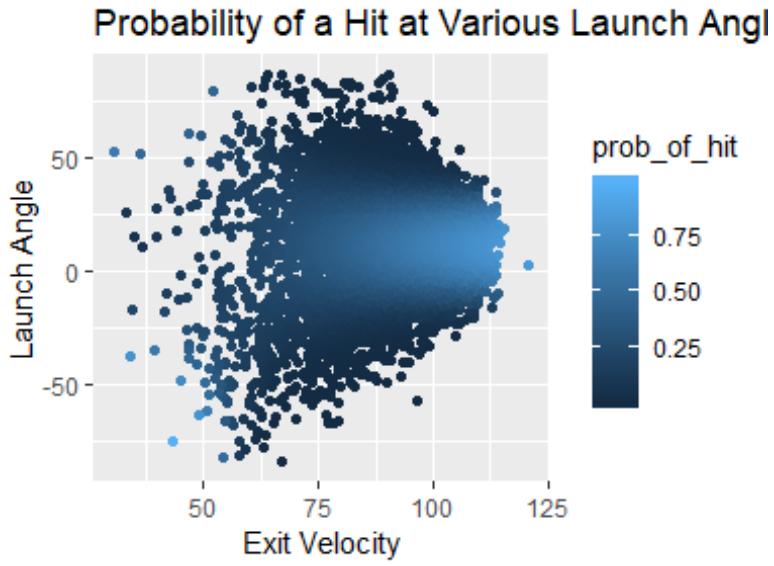
It was also theorized of a posible interaction between exit velocity and launch angle, since players are swinging on a roughly level plane, meaning batted balls with higher launch angles were less likely to be "barrelled", or hit with the middle of the bat (Appendix Model 4).

A quasibinomial model was chosen as the final model, since the overdispersion parameter was 71.6. entered exit velocity and z-score launch angle and squared z-score launch angle used for ease of interpretation (Summary in Appendix).

Results

```
##                    2.5 %    97.5 %
## (Intercept)   0.59966500 2.5081471
## cexit         1.00492727 1.1064558
## zangle        0.25174984 1.4934431
## zangle2       0.04330841 0.4768714
## cexit:zangle  0.93398050 1.0535538
## cexit:zangle2 0.89985840 0.9938723
```

Exit velocity significantly increased the odds of a batted ball being a hit. Squared launch angle was also a significant predictor of batted ball outcome, as the odds of a hit decreased with increased launch angle. The interaction between exit velocity and squared launch angle was also significant; the odds of a hit decreased when batted balls with high launch angles were hit with high exit velocity.



The results are consistent with observations and hitting strategy. Hitting a ball harder should increase the odds of a base hit, since the ball will travel farther than a softer hit ball if they have the same launch angle. Harder hit balls are also harder for a fielder to judge, so they are more likely to drop in for hits. Launch angle has a quadratic effect on the odds of a base hit. Batted balls hit into the ground do not travel far, and pop-ups, batted balls hit straight up are likely to hang in the air long enough for a fielder to get under the ball to make the catch and record an out.

Limitations of the research are that fielder positioning is unknown beyond standard and shifts, which while valuable, isn't reliable as not every team shifts, and each shift looks different depending on the batter and fielding team. Future research could address shifting as well as a player's ability to direct their hits to different sides of the field. Future research could also look at bat speed and bat path, since each players' swing is unique.

References

Gray, R. (2018). Comparing cueing and constraints interventions for increasing launch angle in baseball batting. Sport, Exercise, and Performance Psychology, 7(3), 318.

Hecht, H., & Bertamini, M. (2000). Understanding projectile acceleration. Journal of Experimental Psychology: Human Perception and Performance, 26(2), 730.

Bailey, S. R. (2017). Forecasting batting averages in MLB.

Sievert, C., & Mills, B. M. (2017). Using publicly available baseball data to measure and evaluate pitching performance. In Handbook of Statistical Methods and Analyses in Sports (pp. 55-82). Chapman and Hall/CRC.

https://fivethirtyeight.com/features/the-new-science-of-hitting/

https://www.washingtonpost.com/graphics/sports/mlb-launch-angles-story/?utm_term=.a18cb0db0228

https://www.si.com/mlb/2018/03/21/evolution-swing-home-run-opening-day

Appendix

Final Model - quasibinomial regression model with centered exit velocity, z-score launch angle, squared z-score launch angle, and the interactions between centered exit velocity and z-score launch angle and exit velocity and square z-score launch angle.

```
##
## Call:
## glm(formula = hit ~ cexit * zangle + cexit * zangle2, family =
quasibinomial,
##     data = batted)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7811  -0.8270  -0.3681   0.9142   5.0672
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.20057    0.36160   0.555   0.5791
## cexit          0.05059    0.02419   2.091   0.0365 *
## zangle        -0.43815    0.44356  -0.988   0.3233
## zangle2       -1.76585    0.60804  -2.904   0.0037 **
## cexit:zangle  -0.01275    0.02838  -0.449   0.6533
## cexit:zangle2 -0.05502    0.02353  -2.338   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 71.62476)
##
##     Null deviance: 7521.6  on 5802  degrees of freedom
## Residual deviance: 5844.4  on 5797  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Model 1 - logistic regression with exit velocity as explanatory variable

```
##
## Call:
## glm(formula = hit ~ launch_speed, family = binomial, data = batted)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4953  -0.9345  -0.7155   1.1498   2.7110
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.490235   0.232874  -23.58   <2e-16 ***
## launch_speed  0.053832   0.002512   21.43   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7521.6  on 5802  degrees of freedom
## Residual deviance: 6979.7  on 5801  degrees of freedom
## AIC: 6983.7
##
## Number of Fisher Scoring iterations: 3
```

Model 2 - logistic regression model with launch angle as explanatory variable.

```
##
## Call:
## glm(formula = hit ~ launch_angle, family = binomial, data = batted)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1607  -0.9525  -0.8717   1.4256   1.5865
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.531640   0.030820 -17.250  < 2e-16 ***
## launch_angle -0.005860   0.001017  -5.761 8.37e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7521.6  on 5802  degrees of freedom
## Residual deviance: 7488.2  on 5801  degrees of freedom
## AIC: 7492.2
##
## Number of Fisher Scoring iterations: 4
```

Model 3 - logistic regression model with launch angle and squared launch angles as explanatory variables.

```
##
## Call:
## glm(formula = hit ~ launch_angle + la2, family = binomial, data = batted)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2921  -0.9630  -0.4237   1.0848   5.1942
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   8.111e-02  3.980e-02   2.038   0.0416 *
## launch_angle  3.659e-02  2.374e-03  15.411   <2e-16 ***
## la2          -1.814e-03  7.697e-05 -23.564   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7521.6  on 5802  degrees of freedom
## Residual deviance: 6327.7  on 5800  degrees of freedom
## AIC: 6333.7
##
## Number of Fisher Scoring iterations: 6
```

Model 4 - logistic regression model with exit velocity, launch angle, squared launch angle, and the interaction of exit velocity with launch angle and exit velocity with squared launch angle.

```
##
## Call:
## glm(formula = hit ~ launch_speed * launch_angle + launch_speed *
##     la2, family = quasibinomial, data = batted)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7811  -0.8270  -0.3681   0.9142   5.0672
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -3.783e+00  2.442e+00  -1.549   0.1215
## launch_speed                4.192e-02  2.663e-02   1.574   0.1155
## launch_angle               -9.498e-02  8.005e-02  -1.187   0.2354
## la2                         4.138e-03  2.272e-03   1.821   0.0686 .
## launch_speed:launch_angle   1.658e-03  1.042e-03   1.592   0.1115
## launch_speed:la2           -7.303e-05  3.124e-05  -2.338   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 71.62476)
##
##     Null deviance: 7521.6  on 5802  degrees of freedom
## Residual deviance: 5844.4  on 5797  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6

##
## Call:
## glm(formula = hit ~ cexit * zangle + cexit * zangle2, family =
quasibinomial,
##     data = batted)
##
## Deviance Residuals:
```

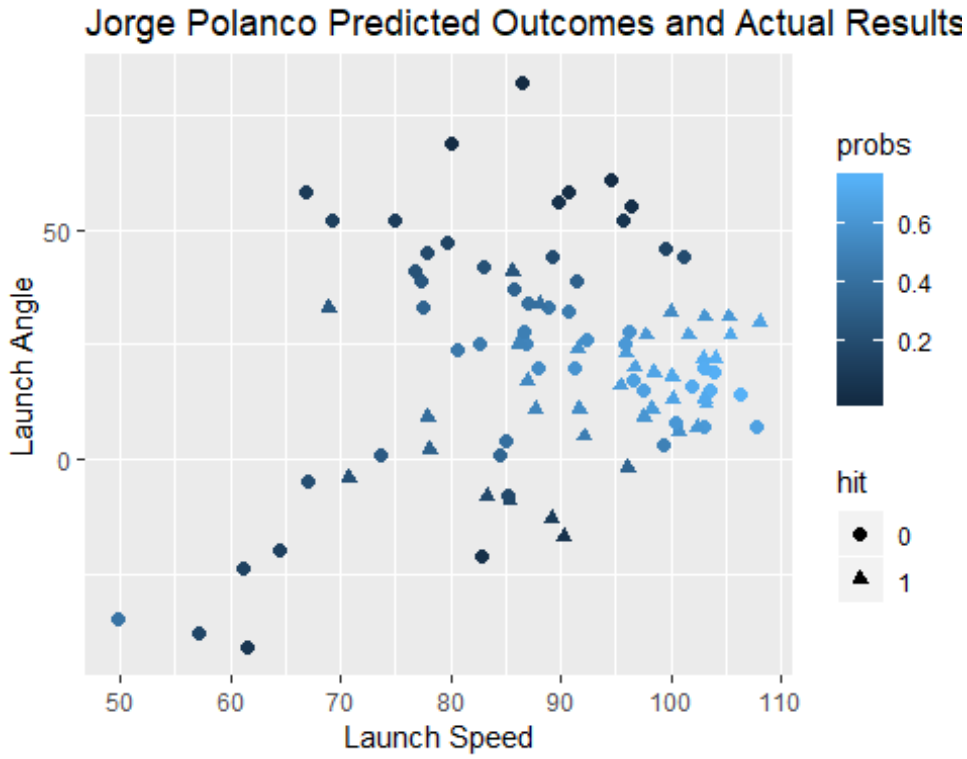
```
##     Min       1Q   Median       3Q      Max
## -1.7811  -0.8270  -0.3681   0.9142   5.0672
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.20057    0.36160   0.555   0.5791
## cexit          0.05059    0.02419   2.091   0.0365 *
## zangle        -0.43815    0.44356  -0.988   0.3233
## zangle2       -1.76585    0.60804  -2.904   0.0037 **
## cexit:zangle  -0.01275    0.02838  -0.449   0.6533
## cexit:zangle2 -0.05502    0.02353  -2.338   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 71.62476)
##
##     Null deviance: 7521.6  on 5802  degrees of freedom
## Residual deviance: 5844.4  on 5797  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Drop in deviance test comparing the significance of the interactions between exit velocity and launch angle, and exit velocity and sqaured launch angle.

```
## Analysis of Deviance Table
##
## Model 1: hit ~ launch_speed * launch_angle + launch_speed * la2
## Model 2: hit ~ launch_speed + launch_angle + la2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5797     5844.4
## 2      5799     6204.5 -2  -360.13  0.08094 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can use the model to assess how lucky or unlucky a batter has been. Jorge Polanco will be used as an example. His batted balls between April 7 and May 15 were downloaded from the Statcast Website and used as a test dataset. The same data cleaning processes were applied to the Polanco dataset, and then the predict function in R was used with the model

and the Polanco dataset.




```
##    preds
## hit  0  1
##   0 45 16
##   1 13 26

## [1] 0.29
```

Of the 100 batted balls in his sample, 29% were missclassified, with a nearly equal amount of lucky hits and unlucky outs. This is a high error rate, however, baseball is a sport where luck is regularly required and sample sizes of 100 can be considered small.