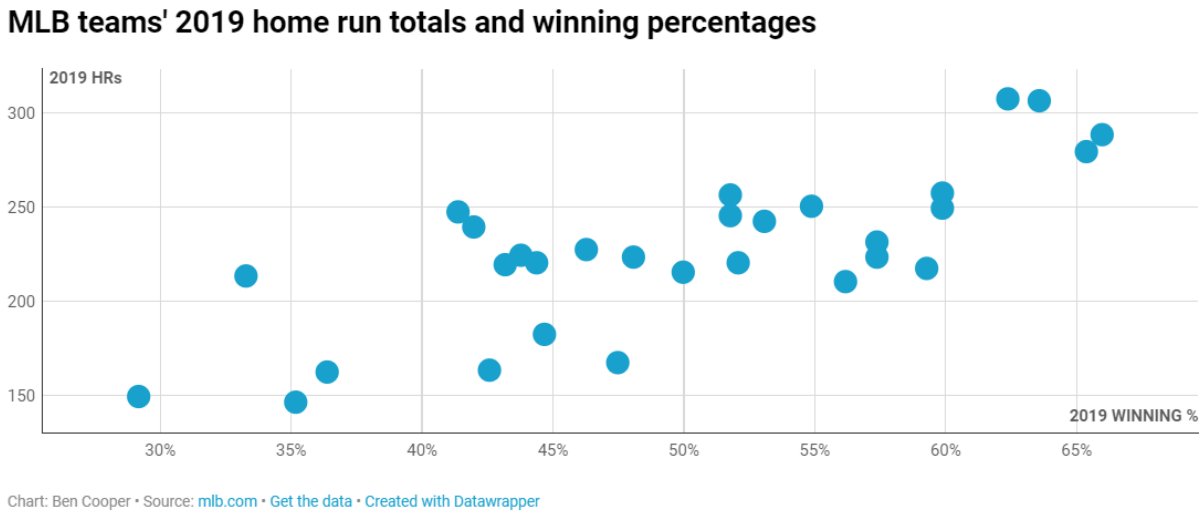
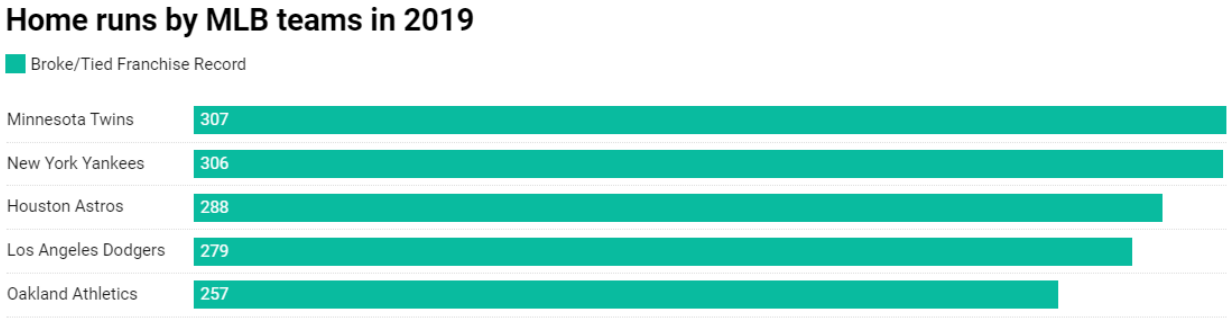
Daniel Allen
Ravi Bhagwat
Dhruvil Dagli

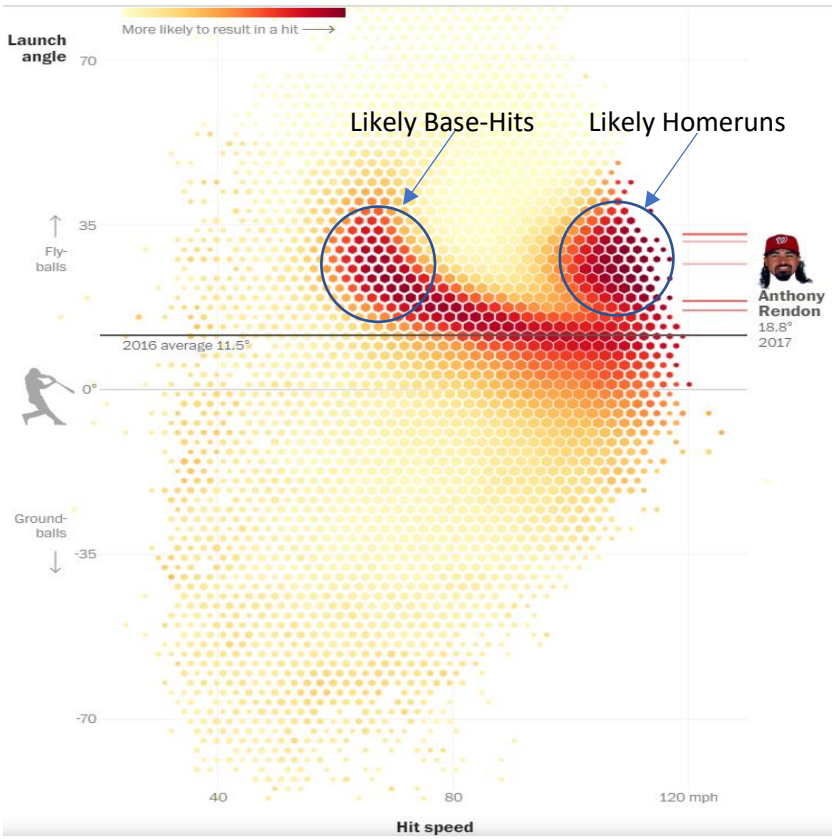
Using Pitcher Statistics to Predict Power Hitting

Major league baseball shattered the previous homerun record in 2019 with 6,776 homeruns, a mark that was 11% higher than the previous record, and 21% higher than the total in 2018. As the MLB progresses forward, prioritizing the homerun has become common place for most major league offenses. This trend is not new, however. Homeruns have been steadily increasing since the origins of professional baseball, and have hit a sharp incline in the last 5 years alongside Statcast data tracking. The desire for the longball is not unfounded however. Homeruns have a significant impact on runs scored, and are correlated with higher winning percentages among teams that hit more homeruns.





In 2019, winning percentage and homeruns have a positive correlation. The top 4 teams in homerun totals also represent the top 4 winning percentages. Of the top 5 teams in homerun totals, all 5 recorded a franchise record. All 5 of these teams also finished in the top 6 teams in win totals. Teams have uncovered the optimal situations to produce homeruns via Statcast tracking. With this focused approach MLB teams and fan are able to visualize when homeruns



happen, along with hit probabilities when studying metrics of each individual bat-ball metric. By looking at exit velocity and launch angle, we can visualize when hits and homeruns are more likely. Players like Anthony Rendon have had success in improving their batting average merely by increasing their launch angle while maintaining a exit velocity above 90mph. From 2016 to 2019, Rendon increased his launch angle by nearly 3 degrees. This resulted in a jump from 20 HR to 34 HR, and an 18% increase in batting average. (His exit velocity decreased by .7 mph during this time) This adjustment by Rendon resulted in an All-star nomination, a Silver Slugger award, and a third place finish in MVP voting. Rendon is a microcosm of the movement towards more homeruns, but is a prime example of how launch angle can positively affect MLB offenses.

While offenses are prioritizing power hitting, pitching staffs face the challenge of stifling power hungry hitters. In this analysis, raw pitching attributes along with basic batter contact variables will be analyzed to gauge past and future pitcher success. By focusing on pitcher arsenals, a linear model with significant predictors of power hitting will be interpreted back to an administration level to analyze the value and trends of players. Ideally, the model we allow pitchers to weaponize their attributes to decrease power hitting and inform front offices of areas of opportunity.

Introducing Regression

By introducing a linear model, we can use regression to estimate the relationship of homeruns on the dependent y value of “Wins”. By choosing Homeruns hit by each MLB offense and Homeruns allowed by their respective pitching staffs, we can see how Homeruns on both sides of the game explain the variance in team wins.

```
lm(formula = W ~ HR + HRAgnst, data = team)

Residuals:
    Min      1Q  Median      3Q     Max
-18.631  -6.141   1.925   5.255  11.794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 80.96696   16.74817   4.834 4.75e-05 ***
HR           0.24551    0.03720   6.599 4.43e-07 ***
HRAgnst     -0.24551    0.05447  -4.507 0.000115 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.037 on 27 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7621,   Adjusted R-squared:  0.7445
F-statistic: 43.25 on 2 and 27 DF,  p-value: 3.808e-09
```

As explained by the output, HR and HRAgainst explain 76% of the variance in wins and are both significant predictors of Wins. The same can be done to predict homeruns, but this time with exit velocity and launch angle.

```
Lm(formula = HR_AB ~ exit_velocity_avg + launch_angle_avg, data = homeruns)

Residuals:
      Min        1Q    Median        3Q       Max
-0.032437 -0.007319 -0.000521  0.008341  0.036633

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.4842716  0.0469571 -10.313  < 2e-16 ***
exit_velocity_avg  0.0056778  0.0005254  10.806  < 2e-16 ***
launch_angle_avg   0.0018454  0.0002614   7.059 8.58e-11 ***
---

Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0124 on 132 degrees of freedom
Multiple R-squared:  0.5703,   Adjusted R-squared:  0.5638
F-statistic: 87.58 on 2 and 132 DF,  p-value: < 2.2e-16
```

Using launch angle and exit velocity to predict homeruns is less adequate of a model however. Only 57% of the variance homeruns are explained what the two metrics. We know exit velocity and launch angle can predict homeruns, and we have evidence of players having success by increasing their launch angle, but the two metrics only explain 57% of the variance in homeruns. By referencing the heat map in the introduction, we can see that there are obvious “sweet spots” in exit velocity and launch angle pairs. In other words, launch angle does not have

a linear relationship with homeruns because of extremely high launch angle hits having a high probability of outs. It is the same story with exit velocity: A hard hit is not predictive of a homerun if it is hit at a low launch angle on the ground. Based on the heat map above, the optimal situation for a homerun seems to be in the 20-35 degree angle range with an exit velocity of over 100 MPH. Rather than deal with this nuance directly, another variable could perhaps be a better indicator of power hitting and pitcher ability.

Expected Slugging Percentage

In this analysis, we will be looking at individual pitcher attributes and their ability to prevent power hitting. However, because launch angle and exit velocity present obvious inadequacies when used separately, we will explain the relationship with power hitting by using Expected Slugging Percentage.

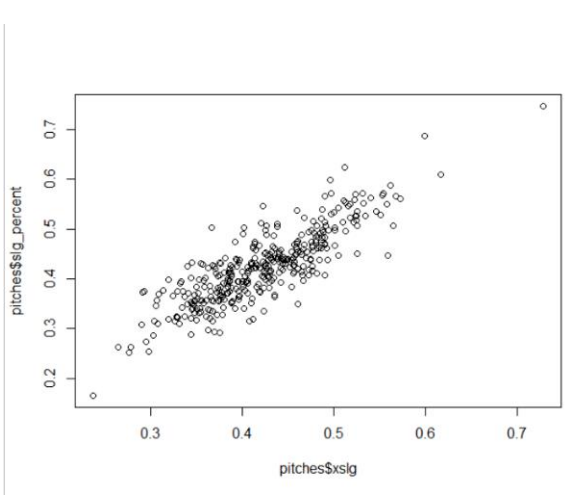
Expecting Slugging Percentage takes into account every bat-to-ball instance and records launch angle and exit velocity. Rather than deal with averages of all hits, we can look at each instance and analyze it separately. We can accomplish this by looking at past bat-to-ball instance and its exit velocity and launch angle have an associated probability for each kind of hit. Traditional slugging percentage is calculated like this:

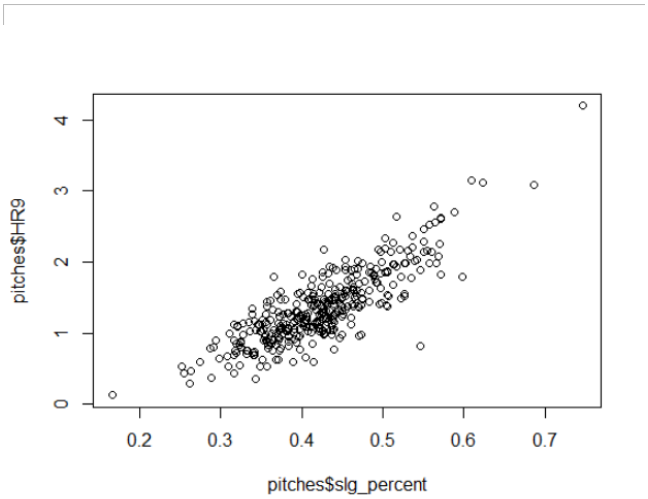
((1B + 2Bx2 + 3Bx3 + HRx4)/AB)

Expected slugging is calculated the same way, but takes each bat-to-ball instance into account and assigns the highest expected probability to a single, double, triple, homerun, or out. Let's say that a 105 MPH, 25 degree angle bat-to-ball instance has the expected probabilities:

((25%:1B) (25%:2B) (10%:3B) (40%:HR) (0%:Out))

HR has the highest probability based on identical exit velocity and launch angle instances, so the traditional slugging metric is updated with a HR.





By visualizing the correlation between Expected Slugging % and Actual Slugging %, we can see how well using expected bat-to-ball statistics can predict slugging percentage. Separate launch angle and exit-velocities do an adequate job at predicting slugging percentage according to the strong correlation between the two statistics from the first plot. From the second scatter

plot, we can see how traditional slugging percentage allowed is associated with homeruns allowed. There is an obvious strong positive correlation in slugging percentage and homerun frequency due to the higher weight attributed to homeruns when calculating slugging percentage.

```
lm(formula = W ~ SLG + SLGAgnst, data = team)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0742 -4.3320  0.1676  4.0572  8.3865

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    70.05      26.73   2.620   0.0142 *
SLG           314.37      37.30   8.428 4.86e-09 ***
SLGAgnst     -289.37      36.77  -7.869 1.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.522 on 27 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8877,  Adjusted R-squared:  0.8794
F-statistic: 106.7 on 2 and 27 DF,  p-value: 1.518e-13
```

Slugging Percentage on both offense and defense also does a better job at explaining the variance in team wins than homeruns, with 88% of the variance in Wins explained by offensive slugging and slugging allowed. The negative estimate for Slugging Against means that a higher slugging percentage allowed results in less wins. For Slugging-for, the positive estimate means that higher offensive slugging results in higher wins.

This provides justification that Expected Slugging will be a formidable variable to analyze in replacement of launch angle and exit velocity, and homeruns specifically. After all, while homeruns are becoming more sought after by MLB offenses, what we are really interested is seeing how pitcher attributes can be analyzed to predict optimal launch angle and exit velocity combinations that result in high power hitting averages. Expected Slugging allows us to examine the undefined “sweet spot” in launch angle and exit velocity, and allows regression to be a tool to predict a dependent variable in a linear model. By using expected slugging, we can quantify launch angle and exit velocity together, provide bias to homeruns, and effectively analyze every bat-to-ball instance.

Variables

By looking at 27 variables across 366 pitchers, 5 categories of pitching will be analyzed to create a predictive regression model. Pitch selection, speed, movement, location, and batter contact will be utilized to predict Expected Slugging Percentage.

Pitch Selection
Fastball %
Breaking %
Off-Speed %
Four-seam Fastball %
Slider %
Change-up %
Curve-ball %
Sinker %

Pitch Speed
Fastball Avg MPH
Off-speed Avg MPH
Breaking Avg MPH

Pitch Movement
Four-seam Fastball Horizontal Movement
Four-seam Fastball Vertical Movement
Slider Average Movement
Change-up Average Movement
Curve-ball Average Movement
Sinker Average Movement

Pitch Location
In-zone %
Edge-zone %
Middle-zone %
3-0 Count %
0-2 Count %
First Pitch Strike %

Batter Contact
Zone-Swing-Miss %
Out-Zone Swing-Miss %
Exit Velocity MPH
Launch Angle

Regression

Linear Modeling, residual analysis, analysis of variance, out of sample testing and variable selection will be performed to create a final regression model to explain the variance in Expected Slugging Percentage.

Pre-processing

Pitch types and their respective average breaks represent the majority of the missing values in the dataset. This is due to the difference in arsenal among different pitchers. While there are 5 major types of pitches, any given pitcher will utilize a subset of them. This results in missing values for certain pitch types for the majority of the data. Intuitively, this missing value should be filled in with “0” because they throw this pitch 0% of the time. However, to avoid inadequacies and adding unneeded variance to the data, specifics pitch percentages were removed from the model. For replacement, Offspeed, Fastball, and Breaking percentages are used in lieu of the specific pitch types. The remaining 35 observations with missing values were filled with 0 when the pitch type was not used, and with column mean for the remainder of the missing data.

Full Model

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.185e-01  9.940e-02   2.199  0.02858 *
exit_velocity_avg      8.833e-03  7.642e-04  11.559  < 2e-16 ***
launch_angle_avg       1.562e-03  2.597e-04   6.014 4.65e-09 ***
avg_horizontal_brk     2.182e-04  1.390e-04   1.570  0.11743
avg_vertical_brk       2.546e-04  4.303e-04   0.592  0.55446
sl_avg_break          -1.302e-03  3.222e-04  -4.043 6.53e-05 ***
ch_avg_break          -1.325e-03  4.014e-04  -3.302  0.00106 **
cu_avg_break          -6.428e-04  2.815e-04  -2.284  0.02299 *
si_avg_break           1.530e-03  6.447e-04   2.373  0.01820 *
n_fastball_formatted  -2.016e-04  1.494e-04  -1.350  0.17795
n_breaking_formatted   9.798e-05  1.462e-04   0.670  0.50317
fastball_avg_speed    -1.025e-03  7.175e-04  -1.429  0.15394
breaking_avg_speed    -1.385e-04  4.260e-04  -0.325  0.74534
n_offspeed_formatted   1.311e-04  1.488e-04   0.881  0.37906
offspeed_avg_speed     2.381e-04  4.169e-04   0.571  0.56828
meatball_percent       6.712e-04  1.302e-03   0.516  0.60650
in_zone_percent        2.372e-04  4.819e-04   0.492  0.62293
edge_percent          -1.809e-04  5.966e-04  -0.303  0.76197
f_strike_percent      -2.935e-05  2.891e-04  -0.102  0.91920
z_swing_miss_percent  -2.156e-03  3.302e-04  -6.528 2.40e-10 ***
oz_swing_miss_percent -1.083e-03  1.668e-04  -6.495 2.91e-10 ***
X30_count_percent     -3.939e-04  7.411e-04  -0.532  0.59541
X20_count_percent     -9.840e-04  2.244e-04  -4.385 1.55e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0178 on 343 degrees of freedom
Multiple R-squared:  0.6961,   Adjusted R-squared:  0.6766
F-statistic: 35.71 on 22 and 343 DF,  p-value: < 2.2e-16
```

The full model produces an R-Squared Value of .696, meaning the variables explain 70% of the variance in expected slugging percentage. 9 of the variables are significant.

Variance Inflation Factor

By using the VIF function in the car package, the variables can be used to predict each other. Values above 5 are generally considered to be influential and are predicted by the other variables in the model.

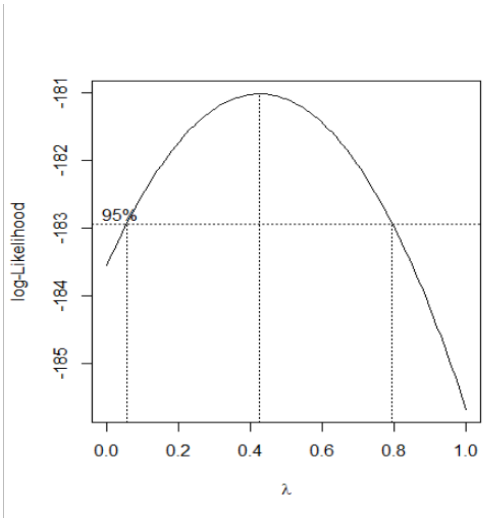
Breaking Pitch % was removed due to a VIF value of over 5.

Cooks Distance

Cooks Distance is used to see if any observations in the data are outliers. Values above 1 are generally considered to be outliers.

There are no values above 1 in the data so no observations will be removed due to high cooks distance values.

BoxCox Transformation



Expected Slugging variable is transformed.

Forward Regression

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.2038871  0.1086926  -1.876 0.061506 .
z_swing_miss_percent  -0.0031391  0.0004560  -6.884 2.68e-11 ***
exit_velocity_avg      0.0131986  0.0010831  12.186  < 2e-16 ***
launch_angle_avg       0.0024067  0.0003337   7.212 3.39e-12 ***
oz_swing_miss_percent -0.0015461  0.0002183  -7.083 7.71e-12 ***
X20_count_percent     -0.0013211  0.0002851  -4.635 5.04e-06 ***
sl_avg_break          -0.0017867  0.0004462  -4.005 7.58e-05 ***
n_fastball_formatted  -0.0004090  0.0001164  -3.514 0.000498 ***
ch_avg_break          -0.0018981  0.0005655  -3.357 0.000874 ***
si_avg_break           0.0025346  0.0008598   2.948 0.003412 **
avg_horizontal_brk     0.0003246  0.0001955   1.660 0.097751 .
cu_avg_break          -0.0008261  0.0003615  -2.285 0.022914 *
fastball_avg_speed    -0.0012172  0.0006816  -1.786 0.074971 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02569 on 353 degrees of freedom
Multiple R-squared:  0.6938,      Adjusted R-squared:  0.6834
F-statistic: 66.64 on 12 and 353 DF,  p-value: < 2.2e-16
```

Backward Regression

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.2038871  0.1086926  -1.876 0.061506 .
exit_velocity_avg     0.0131986  0.0010831  12.186  < 2e-16 ***
launch_angle_avg      0.0024067  0.0003337   7.212 3.39e-12 ***
avg_horizontal_brk    0.0003246  0.0001955   1.660 0.097751 .
sl_avg_break         -0.0017867  0.0004462  -4.005 7.58e-05 ***
ch_avg_break         -0.0018981  0.0005655  -3.357 0.000874 ***
cu_avg_break         -0.0008261  0.0003615  -2.285 0.022914 *
si_avg_break          0.0025346  0.0008598   2.948 0.003412 **
n_fastball_formatted -0.0004090  0.0001164  -3.514 0.000498 ***
fastball_avg_speed   -0.0012172  0.0006816  -1.786 0.074971 .
z_swing_miss_percent -0.0031391  0.0004560  -6.884 2.68e-11 ***
oz_swing_miss_percent -0.0015461  0.0002183  -7.083 7.71e-12 ***
X20_count_percent    -0.0013211  0.0002851  -4.635 5.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02569 on 353 degrees of freedom
Multiple R-squared:  0.6938,      Adjusted R-squared:  0.6834
F-statistic: 66.64 on 12 and 353 DF,  p-value: < 2.2e-16
```

“Both” Regression

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.2038871  0.1086926  -1.876 0.061506 .
z_swing_miss_percent -0.0031391  0.0004560  -6.884 2.68e-11 ***
exit_velocity_avg     0.0131986  0.0010831  12.186  < 2e-16 ***
launch_angle_avg      0.0024067  0.0003337   7.212 3.39e-12 ***
oz_swing_miss_percent -0.0015461  0.0002183  -7.083 7.71e-12 ***
X20_count_percent    -0.0013211  0.0002851  -4.635 5.04e-06 ***
sl_avg_break         -0.0017867  0.0004462  -4.005 7.58e-05 ***
n_fastball_formatted -0.0004090  0.0001164  -3.514 0.000498 ***
ch_avg_break         -0.0018981  0.0005655  -3.357 0.000874 ***
si_avg_break          0.0025346  0.0008598   2.948 0.003412 **
avg_horizontal_brk    0.0003246  0.0001955   1.660 0.097751 .
cu_avg_break         -0.0008261  0.0003615  -2.285 0.022914 *
fastball_avg_speed   -0.0012172  0.0006816  -1.786 0.074971 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02569 on 353 degrees of freedom
Multiple R-squared:  0.6938,      Adjusted R-squared:  0.6834
F-statistic: 66.64 on 12 and 353 DF,  p-value: < 2.2e-16
```

Reg Subsets

All step regression produced the same 12 variables. By taking these 12 variables, we will use subsets to find the top models of different size combination.

```
exit_velocity_avg
launch_angle_avg
z_swing_miss_percent
oz_swing_miss_percent
n_fastball_formatted
sl_avg_break
cu_avg_break
ch_avg_break
si_avg_break
fastball_avg_speed
X20_count_percent
avg_horizontal_brk
```

Comparing Model Statistics

	avg_horizontal_	sl_avg_brea	ch_avg_bre	cu_avg_br	si_avg_bre	n_fastball_	fastball_avg	z_swing_miss_	oz_swing_m	X20_count_	exit_velocity_a	launch_angle_a	bic	cp	rss	rsq	adjr2
5								*	*	*	*	*	-333.53	66.65474	0.277552	0.63507	0.630001
5		*						*	*		*	*	-332.7	67.61015	0.278182	0.634241	0.629161
5						*		*	*		*	*	-329.798	70.96591	0.280396	0.631329	0.626209
5				*				*	*		*	*	-323.374	78.49111	0.285361	0.624801	0.61959
6		*						*	*	*	*	*	-347.836	46.05739	0.262642	0.654673	0.648902
6						*		*	*	*	*	*	-343.815	50.45522	0.265543	0.650858	0.645023
6		*				*		*	*		*	*	-342.669	51.71733	0.266376	0.649763	0.64391
6				*				*	*	*	*	*	-338.402	56.45099	0.2695	0.645657	0.639734
7		*				*		*	*	*	*	*	-357.334	31.65601	0.25182	0.668902	0.662428
7				*		*		*	*	*	*	*	-349.485	39.929	0.257279	0.661725	0.655111
7		*	*					*	*	*	*	*	-349.23	40.20098	0.257458	0.661489	0.65487
7		*					*	*	*	*	*	*	-348.417	41.06823	0.25803	0.660737	0.654103
8		*	*			*		*	*	*	*	*	-358.405	26.45284	0.247067	0.675151	0.667871
8		*				*	*	*	*	*	*	*	-357.167	27.72135	0.247904	0.674051	0.666746
8		*		*		*		*	*	*	*	*	-356.594	28.30972	0.248292	0.67354	0.666225
8	*	*				*		*	*	*	*	*	-356.477	28.42994	0.248372	0.673436	0.666118

The three highlighted models are the final models that will be compared using out of sample testing. 1 model from each size 6, 7, and 8 are chosen based on their stats compared to the other models of their class. The model of size 6 was chosen for its small size and ability to produce the lowest cp, rss, bs, and highest R squared values in its class. In 7, it was by far the best model across all 5 statistics. The 8 variable model performed comparatively to the rest of the 8 variable models but had the best statistics overall. All 3 models PRESS stats are within .01 of each other. Out of sample testing will allow the variables to be cross checked for significance.

Analysis of Variance

```
Response: xslg
                      Df   Sum Sq  Mean Sq  F value    Pr(>F)
sl_avg_break           1 0.048157 0.048157  69.4940 1.666e-15 ***
n_fastball_formatted   1 0.000057 0.000057   0.0828    0.7738
X20_count_percent      1 0.077763 0.077763 112.2166 < 2.2e-16 ***
z_swing_miss_percent   1 0.139794 0.139794 201.7317 < 2.2e-16 ***
oz_swing_miss_percent  1 0.056633 0.056633  81.7254 < 2.2e-16 ***
exit_velocity_avg      1 0.137479 0.137479 198.3912 < 2.2e-16 ***
launch_angle_avg       1 0.041304 0.041304  59.6043 1.171e-13 ***
Residuals            358 0.248083 0.000693
```

After performing anova on the models, it was found the Fastball % has an insignificant variance and should be removed from any models. This means that many pitchers in baseball throw fastballs at a similar rate and the variable lacks variance to be considered a predictor of Expected Slugging Percentage. The removal of this variable make model 1 and 2 the same, so going further, only model 2 and 3 will be compared.

Out of sample Testing

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.9980643  0.2043382  -4.884 2.33e-06 ***
sl_avg_break         -0.0022331  0.0008540  -2.615  0.00971 **
ch_avg_break         -0.0020978  0.0013728  -1.528  0.12828
X20_count_percent    -0.0034556  0.0006430  -5.374 2.43e-07 ***
z_swing_miss_percent -0.0046844  0.0009647  -4.856 2.65e-06 ***
oz_swing_miss_percent -0.0019154  0.0004185  -4.577 8.92e-06 ***
exit_velocity_avg     0.0188500  0.0022455   8.395 1.53e-14 ***
launch_angle_avg      0.0045695  0.0006634   6.888 9.80e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03871 on 175 degrees of freedom
Multiple R-squared:  0.6902,      Adjusted R-squared:  0.6778
F-statistic: 55.71 on 7 and 175 DF,  p-value: < 2.2e-16
```

When the models were run on the “train” set that randomly included the first 50% of the observations, only “Change-up Average Break” was insignificant. This could be justification to remove this variable in the final model. PRESS stats remained the same across all models.

Each model built on the 50% of the observations was also used to predict on the other 50%. By doing this, the sum squared errors on each observation is added to produce the error for each model when predicted on the out sample. When doing these predictions, each error calculation was between .28 and .30, with the larger model having slightly less error in both sets..

K-Folds Cross Validation

```
> m2$results
  intercept       RMSE Rsquared        MAE      RMSESD RsquaredSD      MAESD
1      TRUE 0.02702635 0.646872 0.02064408 0.002968174 0.08907713 0.00222857
> m3$results
  intercept      RMSE  Rsquared        MAE      RMSESD RsquaredSD       MAESD
1      TRUE 0.0268599 0.6561918 0.02077866 0.004705748 0.08912241 0.003135492
```

The model with 6 variables holds the lowest MAE, RMSESD, and MAESD, while the 8 variable model has the highest R squared value. Model 2 is deemed the winning combination of variables to predict Expected Slugging Percentage do to the smallest model size and low error despite having an r-squared value less than Model 3. The difference in these 2 models is the addition of “change up break” in the 7 variable model, and it is determined not worth the 1% hike in R-squared value. The removal of “Change-Up Break” is also justified by its insignificance in the out of sample testing.

Final Model

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.2871237  0.0967914  -2.966  0.00321 **
sl_avg_break         -0.0019749  0.0004374  -4.515 8.59e-06 ***
X20_count_percent    -0.0013581  0.0002947  -4.609 5.64e-06 ***
z_swing_miss_percent -0.0036512  0.0004509  -8.098 8.78e-15 ***
oz_swing_miss_percent -0.0014211  0.0002060  -6.898 2.39e-11 ***
exit_velocity_avg     0.0126895  0.0010788  11.763  < 2e-16 ***
launch_angle_avg      0.0027194  0.0003256   8.353 1.47e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02685 on 359 degrees of freedom
Multiple R-squared:  0.6547,      Adjusted R-squared:  0.6489
F-statistic: 113.4 on 6 and 359 DF,  p-value: < 2.2e-16
```

Slider Average Break- Higher break results in smaller XSLG

0-2 count %- Higher percentage of counts pitched to 0-2 result in smaller XSLG

Zone Swing and Miss %- Higher percentage of swings and miss in the zone results in smaller XSLG

Out-Zone Swing Miss %- Higher percentage of swings and miss out of zone result in smaller XSLG

Exit Velocity MPH- Higher average MPH of pitches hit result in higher XSLG

Launch Angle- Average angle of pitches hit result in higher XSLG

Implications

The final model of 6 variables explains 65% of the variance in XSLG. By starting with 22 variables that explained 69% of XSLG, removing 16 only resulted in a decrease of 4% of R-Squared. All variables and their variances are significant and the final model is used to predict XSLG in the original data. After doing this, rank is compared across our original variables in the introduction. Homeruns per 9, Actual Slugging Percentage, Expected Slugging percentage, and Predicted Slugging percentage based on the model are all used to rank each player. Sorting first by XSLG-Model Prediction, the players with the largest difference were subset. From there, the total differential of XSLG-SLG rank and Model Rank-XSLG rank were subset to leave the top and bottom 2 in total differential. (Higher rank means batters have smaller power hitting averages against)

last_name	first_name	slg_percer	HR9	xslg	PredictedX	HR Rank	SLG Rank	XSLG Rank	Pred XSLG	model rank diff	XSLG-SLG rank diff
Burnes	Corbin	0.623	3.122449	0.754907	0.666437	364	364	335	85	-250	-29
Herrera	Kelvin	0.462	1.409002	0.705642	0.673513	218	273	224	107	-117	-49
Perdomo	Luis	0.372	0.75	0.694648	0.737185	33	89	181	320	139	92
Kintzler	Brandon	0.359	0.789474	0.67841	0.743932	38	68	136	338	202	68

For the top 2 green players, their ranks illustrate undervalued players based on our final model. In the case of Corbin Burnes, he is third to last in HR and Actual Slugging, but based on model variables, he is predicted to be 85th in XSLG. This resulted in the highest rank differential of 250. For Kelvin Herrera it is a similar story, our model makes the case that he is very under-valued based on the variables that were measured. For red “overvalued” players it is the opposite. HR and slugging ranks for both players are near the top, while the model ranks them near the bottom. One glaring difference in these undervalued and overvalued pitchers is their homerun totals. The average HR per 9 for the data is 1.35, and in the case of the top two undervalued and overvalued players, their HR9 values are both far from the mean. This goes back to the importance of homeruns and their ability to make or break a pitcher. This becomes especially true for relief pitchers who pitch less innings, and one mistake is reflected in their per 9 inning average. In the case of Corbin Burnes, if exit velocity and launch angle are looked at separately, his launch angle is below the average and exit velocity is only 1 MPH over the average. Intuitively, this doesn’t coincide with what we know about launch angle, but it highlights the weakness of looking at launch angle averages as a whole rather than analyzing each instance. It is clear that Corbin Burnes is susceptible to mistakes that lead to homeruns, because the model is unable to reflect his HR allowed rank of 364th. For Brandon Kintzler, his Slugging differential suggests he is a pitcher of good fortune. The model also predicts him even worse than his expected numbers, placing him at 338th. By referencing his HR9 numbers, his high expected slugging and predicted expected slugging ranks can be explained by his ability to prevent homeruns. Compared to the average, Kintzler does a great job at preventing homeruns. By looking at other variables not included in the analysis, Kintzler has the 5th highest sinker rate, but ranks in the bottom 40 for zone and out of zone swing and miss. This coupled with a launch angle below the average can be interpreted that Brandon Kintzler is a pitcher who induces contact in the form of ground balls due to the heavy reliance on his sinker. Because the model did not include sinker variables, his tendency to induce contact was measured as a negative effect on Expected Slugging percentage when in reality, his ability to induce ground balls is used to prevent homeruns and high slugging percentage.

last_name	first_name	slg_percer	HR9	xslg	PredictedXS	HR Rank	SLG Rank	XSLG Rank	Pred XSLG Rank	Model diff	XSLG-SLG diff
Hill	Rich	0.395	1.546392	0.630919	0.6119999	260	128	27	3	-24	-101
Crick	Kyle	0.432	1.836735	0.640342	0.6344241	304	212	41	21	-20	-171
Dyson	Sam	0.357	0.869565	0.661616	0.7078459	50	62	88	232	144	26
Bradley	Archie	0.371	0.632022	0.66972	0.7102217	17	84	110	240	130	26

Selecting players by placing bias on their Predicted rank will allow us to focus on undervalued and overvalued pitchers based on the model. In the case of Rich Hill, the stark contrast of his HR rank and Predicted XSLG Slugging rank can provide justification that he is undervalued. Again, HR9 seems to be ignored in our model, but his Expected Slugging compared to his actual rate suggests he is a victim of tough luck, especially in the HR category. If his Expected Slugging and Predicted Expected Slugging are

combined in analysis, a case can be made that Rich Hill possess the arsenal that should make him a better pitcher than what his XSLG values represent, and that his slugging values are inflated due to naturally occurring hard luck.

Conclusions

The strength of expected slugging percentage is that it allows the analyst to look at pitchers in a vacuum. Often times baseball is a random game, and pitchers and hitters can benefit or suffer from where the ball lands. While slugging measures the actual events that occurred, expected slugging analyzes each hit and quantifies what *should* have occurred. In the case of Rich Hill, homeruns seem to cripple his actual slugging numbers, while Expected Slugging and model predictions give him the benefit in his individually allowed bat to ball metrics. Many things could justify this. Rich Hill could be allowing many deep fly balls that are barely sneaking over the wall that could have resulted in an out if it were marginally shorter. He also could be allowing many line-drive doubles down the line or in the gap. One likely explanation is that Rich Hill pitched his 2019 season in the NL West, where 2 of the most consistently power hitting friendly ball parks reside: Coor's Field in Colorado and Chase Field in Arizona. Many different known and unknown variables can explain the randomness in baseball. The model created in this analysis is meant to remove randomness in order to look at things more objectively. Variable selection was an important task because of the unwanted bias and influence certain variables could cause. To avoid bias, measurements were chosen in lieu of calculations and averages. For instance, "Hard Hit Percentage" could have been included in a model to explain Expected Slugging, but was ruled out due to what is known about how Expected Slugging is calculated. Launch angle averages and exit velocity averages were included instead, in order to maintain the goal of predicting a power hitting metric using objective measurements. In the end, the model favored batter-dependent variables in swing and miss percents and exit velocity/launch angle. However, "Slider Average Break" and "0-2 Count Percent" were included and are in total control of the pitcher. Ideally, the final model would be entirely composed of variables that were in the pitcher's control in order to maintain objectivity and provide insights on pitcher's arsenals to be used for scouting and development.

From the standpoint of a pitcher's arsenal, we know that **slider break and getting ahead in the count are significant in decreasing expected slugging percentage.** Overall, we are left with a model that ventures a step further than Expected Slugging percentage, and provides bias to what happens during the pitch up to the moment the ball hits the bat. Regression was useful in predicted the y value of expected slugging percentage and hopefully in the future, a model can be perfected to provide a data driven insight into players like Rich Hill or Brandon Kintzler. Front offices, scouting departments, and fans at home can benefit from looking at how pitchers metrics in the stat cast era translate to wins. Regression is the tool that makes this possible.