Predicting Baseball Performance

Undergraduates: Malik Scott, Jessica Ho, Jack Lichtenstein & Amber Potter
Project Leads: Phuc Nguyen, Sicong Liu, Greg Appelbaum, Tim Sell & Marc Richard

Overview

This project aims to determine if assessment data collected from baseball players participating in the USA Baseball Prospect Development Pipeline (PDP) predicts future batting performance. Assessment data includes measures of visual skills and physical abilities. These are compared to collegiate game statistics or batting propensity data through hierarchical regression analyses to inform scouts about the likely production of these developmental prospects. The final product is an application that uses an athlete's assessment results to produce performance summary graphs for the individual, compared to other athletes and inferential models for the relationships between assessments and performance.

Data

Assessments used as Predictor Variables:

- Player demographics: age, height, weight, position
- PDP athletic measurements: general athleticism/agility
- RightEye vision assessments: vision and processing

Performance Statistics used as Outcome Variables:

- College NCAA/CCBL game statistics: AVG, OPS, Weighted OBP, Strikeout Rate, Walk Rate
- Trackman pitch-level data: Z-Swing Propensity, O-Swing Propensity, Contact Percent, Average Launch Angle, Average Distance, Average Hit Velocity

Modeling

Fit separate models for each outcome variable described in modeling results table below, separately for college stats and trackman datasets

Hierarchical Linear Regression:

- Used to observe whether adding variables significantly improves a model's ability to predict outcome variables

Modeling Steps

Outcome ~ 1
Outcome ~ 1 + Demographics
Outcome ~ 1 + PDP
Outcome ~ 1 + RE
Outcome ~ 1 + Demographics + PDP
Outcome ~ 1 + Demographics + RE
Outcome ~ 1 + PDP + RE
Outcome ~ 1 + Demographics + PDP + RE

Modeling Results

CollegeStats Outcome Variables

Outcome Variables	Results
AVG	NS
OPS	Significant Baseline Model
Weighted OBP	NS
Strikeout Rate	Significant Baseline Model
Walk Rate	NS

Trackman Outcome Variables

Outcome Variables	Results
Z-Swing	NS
O-Swing	NS
Contact %	NS
Average Launch Angle	Significant Baseline Model
Average Distance	Significant Baseline Model
Average Hit Velocity	Significant Baseline Model

Best Model for Strikeout Rate
Strikeout Rate ~ 1 + Demographics

Term	Estimate	P.Value	Terms	Estimates	P.Values
(Intercept)	−0.091	0.588	Position Stats Dh	0.008	0.794
Position Stats Inf	**−0.036**	**0.007**	**Position Stats Of**	**−0.032**	**0.029**
Position Stats P	−0.040	0.571	Height In	0.004	0.136
Weight	**0.001**	**0.039**	**Ncaa**	**−0.113**	**0.000**

Regularized Regression:

Ridge and Lasso

- Prevent overfitting by shrinking coefficients to zero
- Results did not show additional significance

Measuring Vision and Athleticism

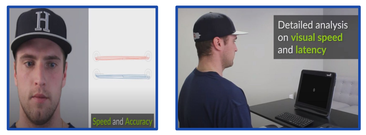
Overview:

PDP athletic assessments measure players' movement and decision making using proximity sensors. Examples of players performing the 30 yard sprint and the Green Box test are shown below and the full list of tests can be found here[1].

RightEye vision tests are quantitative eye-tracking assessments that measure functional vision skills and include; dynamic visual acuity, smooth pursuit and decision making tasks, among others.

Example PDP Assessments

Example RE Assessments



Multiple Imputation:

MICE Predictive Mean Matching

Observations:

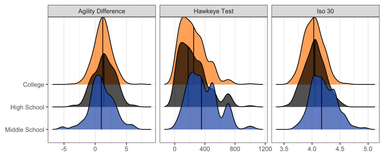
Evaluated data for normalcy, multicollinearity and sensibility. Found good adherence with,

- Differences in PDP scores by age group
- Differences in RightEye scores by year

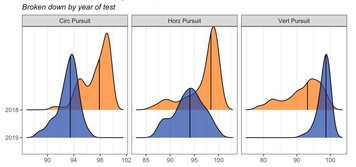
College Stats Summary of Missingness

Variables	Percent Missing	More Variables	More Percent Missing
Green 3	**23.08%**	**Reaction To Go Sec**	**22.56%**
Iso30	**22.56%**	**Green Box**	**22.56%**
Choice Processing	8.72%	Choice Speed	8.72%
Choice Latency	8.72%	Dynamic Visual Acuity	1.03%
Weight	3.08%	Height In	2.05%
Discriminate Processing	8.21%	Discriminate Speed	8.21%
Discriminate Latency	8.21%	Circ Pursuit	4.62%
Horz Pursuit	4.62%	Vert Pursuit	4.62%
Hawkeye	**50.26%**	**Iso10**	**22.05%**
Ethnicity	1.54%	Id	0.00%

College athletes tend to perform better on most athletic assessments
Black lines denote median values



Distribution of smooth pursuit test scores
Broken down by year of test

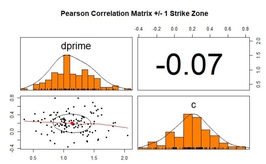


Measuring Batting Performance

Defining a Strike Zone:

- **Signal Detection Theory:** helps differentiate variability d' and c
- Consistent with calls from umpires

Pearson Correlation Matrix +/- 1 Strike Zone



Distribution of called balls and strikes
Boxes denote potential defined strike zones

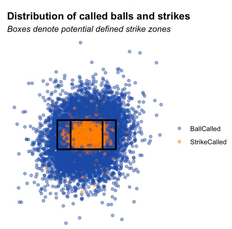
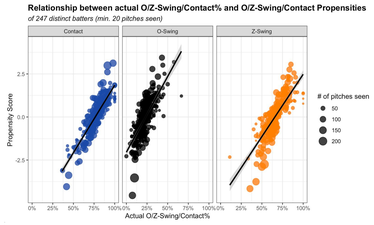


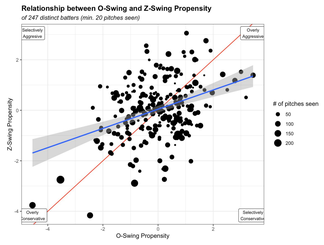
Plate Discipline and Coordination:

- Contact%: Number of pitches on which contact was made / swings
- O-Swing%: Swings at pitches outside zone / pitches outside zone
- Z-Swing%: Swings at pitches inside zone / pitches inside zone

Relationship between actual O/Z-Swing/Contact% and O/Z-Swing/Contact Propensities
of 247 distinct batters (min. 20 pitches seen)



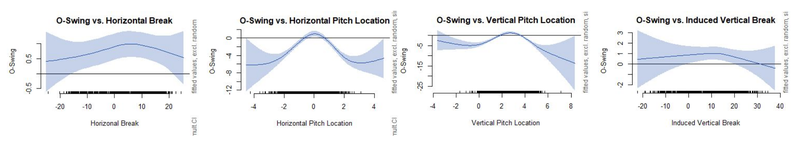
Relationship between O-Swing and Z-Swing Propensity
of 247 distinct batters (min. 20 pitches seen)



Modeling Plate Discipline:

GAMM Models: models Z-Swing, O-Swing, and Contact percentages

- Accounts for varied pitch difficulty and allows for non-linearity



College Performance Statistics:

In addition to the advanced metrics calculated for the trackman data, we are also interested in analyzing game statistics like AVG, OPS, walk and strikeout rates for college players

Compiled Variables form Two Datasets:

College Stats

- 195 players
- Filtered for > 50 AB
- Removed outliers > 3 SDs

Trackman

- 98 players
- Filtered for > 40 pitches faced
- Removed outliers > 3 SDs

Observations and Discussion

Findings:

While several demographic variables predicted batting performance, none of the assessment variables did

Limitations:

- Limited number of observations, included few pitches faced
- Imputation to account for missing data
- Interesting variables missing without ability to impute
- Disparities between variables according to year

Future Works:

- Repeat with more data
- Cross-sectional and longitudinal analyses

Shiny Application

- App can be found here[2] and shows individual compared to other athletes, filtered by age, position, player ID, etc.

Acknowledgements and References

Special thanks to Drew Pomeroy, Matt Pajak, Jules Johnson, and Russell Hartford from USA Baseball and to RightEye for supplying our data and expert domain knowledge.
Thanks to Paul Bendich, Gregory Herschlag, and Ariel Dawn for organizing this program and making this opportunity possible.

[1] If you're viewing this in person, the full link is: https://cloud.3dissue.com/144263/144436/230893/AssessmentDescriptions/index.html?r=21

[2] If you're viewing this in person, the full link is: https://jacklich10.shinyapps.io/USABaseball/