A Whole New Ballgame?

A Statistical Investigation into the MLB Network Top 100 Players List

Honors Thesis Project
Fall 2020

Jonathan Olds

Presentation Outline

01 – Background

02 – History of Baseball Statistics

03 – Data Collection

04 – Methods

05 – Results

06 – Discussion

07 – Final Thoughts

01 - Background

Top 100 Players Lists – By the Numbers

100

Players selected and ranked from 1–100 by MLB Network baseball analysts each spring as the best in the sport for that upcoming season

10

Consecutive years that this list has been released, from 2011 to 2020.

1200

Players that take the field for a major league team in any given season, based on 30 teams with active roster sizes of 40 players.

02 - History of Baseball Statistics

The Subjectivity of ‘Best’

Conventional

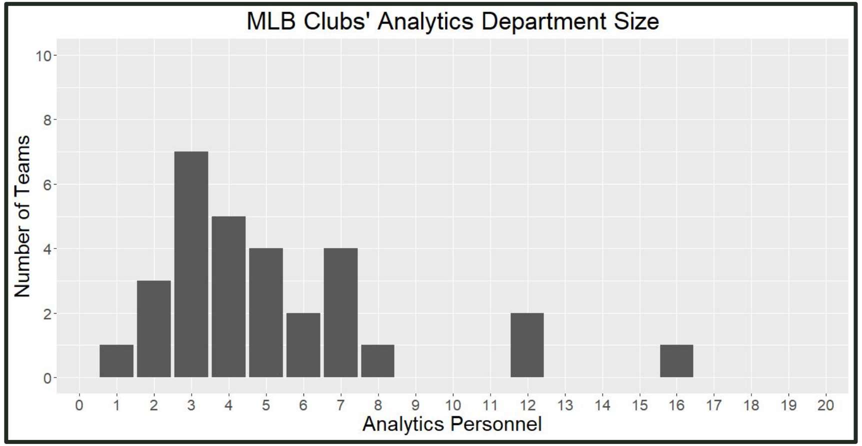
- Present since baseball’s inception
- Henry Chadwick
    - Base hits & batting average
- 1900’s
    - RBI (1907) & ERA (1912)
- Accepted and widely available by mid-twentieth century
- Quicker to calculate, usually easiest to interpret
    - Counting statistics

Sabermetric

- Initial emergence in 1970’s
- Bill James
    - Coined ‘sabermetrics’
    - Baseball Abstract
- *Moneyball*
    - Billy Beane – Oakland A’s
    - Sabermetric approach
- Baseball Reference & Fangraphs
    - Quantity & accessibility
    - wOBA, FIP, WAR
- More involved calculations
    - Account for context

The State of the 30 MLB Teams

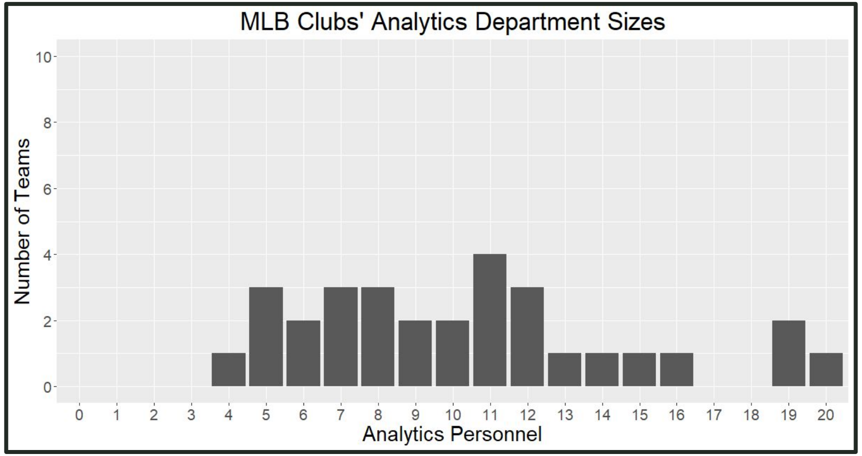
2016



MLB Clubs' Analytics Department Size

5.2
Average

2019



MLB Clubs' Analytics Department Sizes

10.33
Average

The State of Players

In a 2018 poll asking 70 MLB players which statistic they valued the most:

Position Players	31 Conventional	4 Sabermetric
Pitchers	29 Conventional	6 Sabermetric

The State of News Media

A 2017 investigation into baseball articles in the New York Times revealed:

Time Period	Average # of Conventional Statistics	Average # of Sabermetric Statistics
2001–2003 (Pre–*Moneyball* book)	5.94	0.15
2004–2006 (Post–*Moneyball* book)	4.36	0.05
2012–2014 (Post–*Moneyball* film)	5.45	0.41

- Researchers noted a near–threefold increase in the average number of sabermetric statistics mentioned per article in the period immediately following the film release of *Moneyball* as compared to the three year period before the release of the book
- However, conventional statistics still outnumbered sabermetric statistics in absolute terms by a value of 25 to 1.

Do baseball analysts favor conventional statistics or sabermetrics in player evaluation?

03 - Data Collection

Selecting the Lists and Players

Why the Top 100 Lists?

- Scope of the Lists
    - 100 players = broad range of factors

- Independent Analysts
    - Unaffiliated with teams = reduction of bias
        - Represent the group of interest

Selecting the Player Pool

In 2019:

- Only 25% of eligible batters played in half their team's games
    - Limited contribution

Selection Criteria

- Make a Top 100 List OR
- Batters = average 251 or more at-bats over the previous three-year span
- Pitchers = average 41 or more innings-pitched over the previous three-year span
- Rookies

Player Pool – Example

Season	At-Bats	3-Season Rolling Average
2013	240	240
2014	582	411
2015	476	432.667
2016	578	545.333
2017	602	552
2018	574	584.667
2019	489	555

Selecting Player Statistics

Sources

- Baseball Reference
- Fangraphs

Variables

- Conventional/Sabermetric
    - 67 for batters
    - 49 for pitchers
- Awards
    - All-Star
    - Gold Glove & Silver Slugger
    - Most Valuable Player
    - Cy Young
    - Rookie of the Year
- Descriptive
    - Team
    - Age
    - Season
    - Player Identifier

Variable Exclusions

- Irrelevant for player evaluation
    - Strikeouts per win
- Conventional counting statistics that could be replaced with rate statistics
    - Home runs allowed vs. Home runs allowed per 9 innings
- Sabermetric statistics that were not the newest edition
    - Ex: wRC vs. wRC+
- Statistics that offered player value in terms of runs instead of wins
    - Runs Above Average vs. Wins Above Average

Final Dataset Characteristics – Batters

3,357
Player–Seasons

696
Top 100
Player–Seasons

2,661
Non–Top 100
Player Seasons

58
Variables

21
Conventional
Statistics

37
Sabermetric
Statistics

Final Dataset Characteristics – Pitchers

4,774
Player–Seasons

304
Top 100
Player–Seasons

4,470
Non–Top 100
Player Seasons

41
Variables

15
Conventional
Statistics

26
Sabermetric
Statistics

04 - Methods

Logistic Regression (LR) Modeling

Goal: Predict Top 100 List membership for batters and pitchers in separate modeling setups

Basic Model Structure

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

- X_{ik} represents the value of the kth statistic for batter i
- $\beta_1, \beta_2 \ldots \beta_k$ represent the coefficients for each statistic 1,2⋯.k
- Models built with 70% of the data, with 30% used for prediction
- p_i represents the probability of batter i making the top 100 list in a given season
- Use fitted probabilities to assign membership (or non-membership) to the list
- Highly predictive models signify that analysts weight the subset of statistics used in those models more heavily in their player evaluation

Variable Selection Methods

Principal Components Analysis

- Creates linear combinations of the input variables
- Resulting components are uncorrelated
- Selected between 2 and 39 PCA components depending on the model
- Conducted *prior to* regression modeling

LASSO Penalization

- Penalizes, the coefficients corresponding to the input variables
- Forces all coefficients towards zero, and some to zero, thus eliminating their effect
- Penalization depends on a parameter,λ , determined through 10-fold cross validation
- Conducted *during* regression modeling

Variable Selection Methods (Continued)

Principal Components Analysis

PCA was applied to **six** sets of predictor variables:

- Conventional Batters Statistics
- Sabermetric Batters Statistics
- Conventional Pitchers Statistics
- Sabermetric Pitchers Statistics
- Conventional & Sabermetric Batters Statistics
- Conventional & Sabermetric Pitchers Statistics

Components used in **eight** LR models

LASSO Penalization

LASSO Penalization was applied to **six** sets of predictor variables:

- Conventional Batters Statistics
- Sabermetric Batters Statistics
- Conventional Pitchers Statistics
- Sabermetric Pitchers Statistics
- Conventional & Sabermetric Batters Statistics
- Conventional & Sabermetric Pitchers Statistics

Components used in **six** LR models

Prediction Methodology

Comparison of:

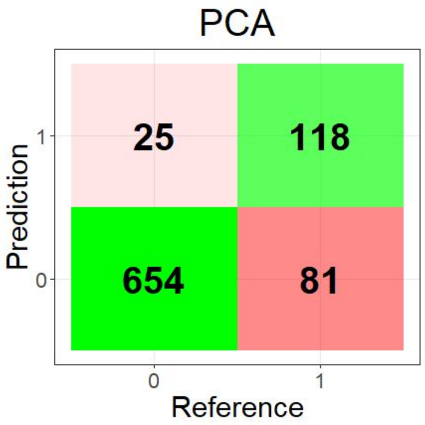
- Predicted Top 100 Players
- Actual Top 100 Players

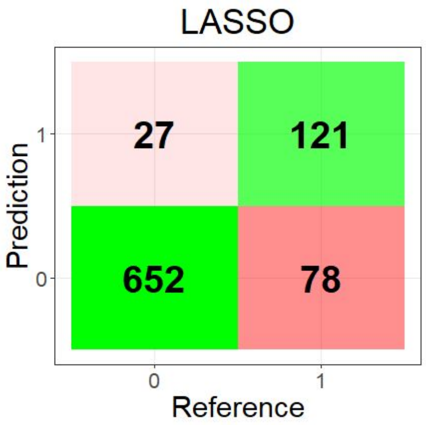
5 Performance Measures:

- Accuracy
- **Sensitivity**
- Specificity
- Positive Pred. Value
- Negative Pred. Value

05 – Results

Confusion Matrices for Conventional Batters Models





- 118 Top 100 players identified correctly
- 654 non–Top–100 players identified correctly
- 25 players incorrectly identified as Top–100 players
- 81 players incorrectly identified as non–Top–100 players

- 121 Top 100 players identified correctly
- 652 non–Top–100 players identified correctly
- 27 players incorrectly identified as Top–100 players
- 78 players incorrectly identified as non–Top–100 players

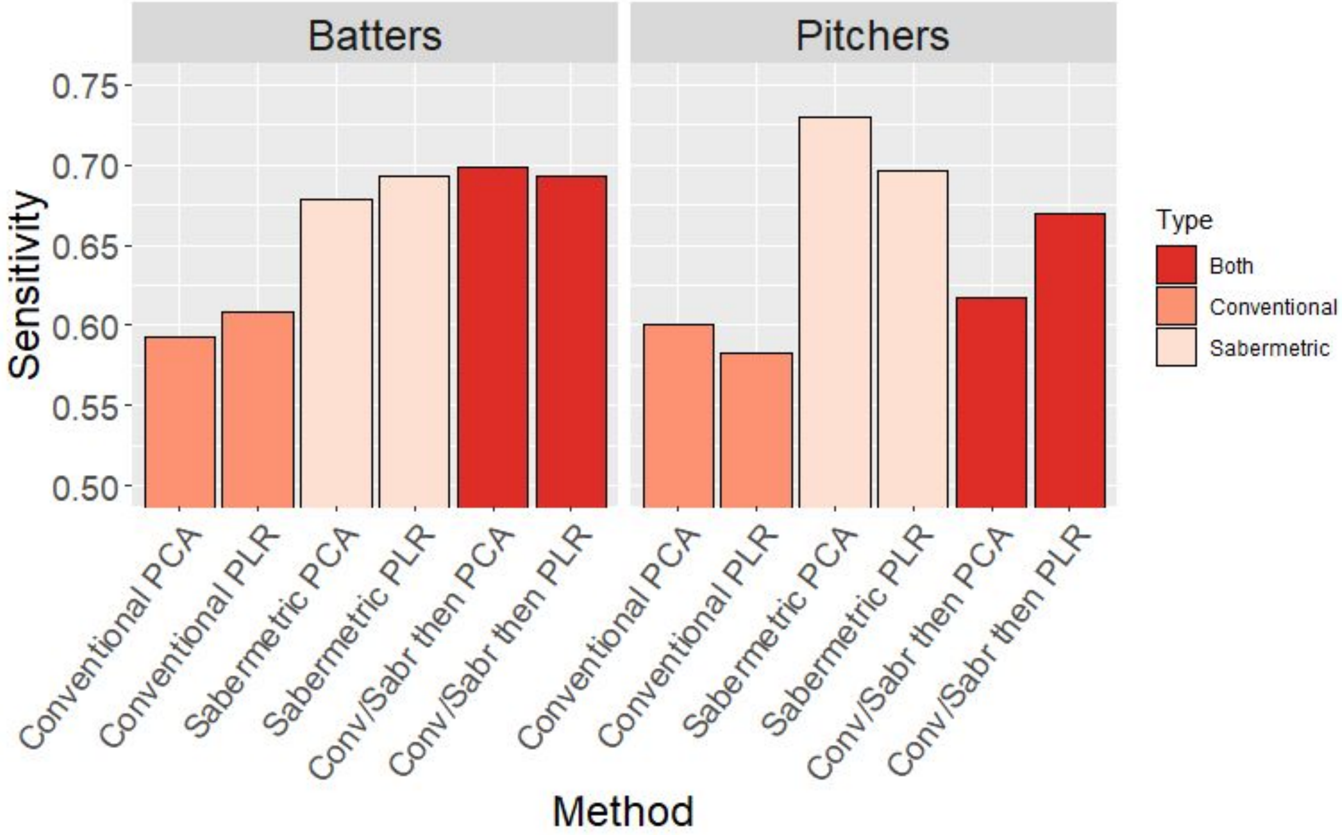
Comparison of Accuracy Measures – PCA Models

Players	Variables	Accuracy	Sensitivity	Specificity	Positive PV	Negative PV
Batters	Conventional	0.879	0.593	0.963	0.825	0.890
Batters	Sabermetric	0.880	0.678	0.940	0.767	0.909
Batters	Conventional & Sabermetric	0.885	0.698	0.938	0.763	0.914
Pitchers	Conventional	0.950	0.600	0.986	0.821	0.959
Pitchers	Sabermetric	0.961	0.730	0.985	0.832	0.972
Pitchers	Conventional & Sabermetric	0.950	0.617	0.985	0.807	0.961

Comparison of Accuracy Measures – LASSO Models

Players	Variables	Accuracy	Sensitivity	Specificity	Positive PV	Negative PV
Batters	Conventional	0.880	0.608	0.960	0.818	0.893
Batters	Sabermetric	0.884	0.693	0.940	0.771	0.913
Batters	Conventional & Sabermetric	0.883	0.693	0.938	0.767	0.913
Pitchers	Conventional	0.947	0.583	0.985	0.798	0.958
Pitchers	Sabermetric	0.954	0.696	0.981	0.792	0.969
Pitchers	Conventional & Sabermetric	0.951	0.670	0.980	0.778	0.966

Comparison of Sensitivity Across Variable Selection Methods



06 – Discussion

Overall Results

Batters/Pitchers

- Batters
  - Sabermetrics preferred over conventional alone
  - Conflicting results as to if conventional statistics have value when used with sabermetrics
- Pitchers
  - Sabermetrics clearly preferred over conventional alone
  - Conventional not preferred even if utilized with sabermetrics

Theories

- Increase in number, accessibility, effectiveness and marketing of sabermetric statistics prior to 2010
- Greater momentum to the movement that conventional statistics are not useful for pitchers
  - Win-Loss Record & ERA in favor of FIP & SIERA
- Conventional batting statistics still interpreted as valuable
  - Home Runs & Walks

Limitations

- Analysis limited to groups of statistics
  - Results not statistic-specific
- Important factors not included
  - Intangibles
  - Defensive Position
- Do analysts use logistic regression?
- Imperfect player pool selection system
  - No requirement to catch all Top 100 players

07 – Final Thoughts

Final Thoughts

Further Research

- Including positional adjustment
  - Potentially different statistics have unique levels of importance
  - 1st basemen vs. 2nd basemen
- Examining which statistics are important for ranking the players
- Trends over time

Big Picture

- Sabermetrics
  - Considerably more valuable among analysts
  - Poised for a larger role in baseball
- Conventional Statistics
  - Not obsolete – analysts may still use them
  - Many non-analytical groups prefer them
- Future
  - Analytics are a spectrum

ROCKIES
SPORTS AUTHORITY
WELLS FARGO
Coca-Cola
Coors LIGHT
KING Soopers
Let's Go Places
TUACA
TUACA
TUACA
TAVERN
Five Years Out
JIMMY JOHN'S
Stranahan's
MAXX HD
WELCOME TO
SINCE 1993

Questions?