Lesson_28_Boardsheet

Kevin Cummiskey

4/3/2020

Review

What characteristics of a hitter does OPS quantify?

On base % + slugging percentage ⇒ correlated with runs production

"gets on base" "hits with power"

Why might one use OPS+ or adjusted OPS+ instead of OPS?

1. Equivalency Coefficients
2. Trajectories

compare hitters of different time periods

takes into account ball park effects

What are some limitations of these statistics?

They don't directly quantify the number of runs the player will contribute.

Linear Weights

Recall we talked about the relationship between Runs and Wins in Chapter 4 of the Marchi text. What was the rule of thumb we used?

10 Runs ≃ 1 Win

(*Understanding Sabermetrics* pg 83) Ideally, we want a statistic for batters that:

- quantifies their contribution to runs scored. ✓
- is not based upon the situations the batters faced when they came to the plate (since their batting actions did not create those situations). ✓

How about RBI's? A batter is awarded Runs Batted In (RBIs) for most situations when his plate appearance results in runs scored. Why are RBIs not a good measure of a player's contribution to runs scored?

It's not good because it depends upon factors not related to the player's hitting. Specifically, the number of runners on-base.

Various researchers have proposed linear models to quantify a player's contribution to runs scored. These models weight individual statistics. For the purposes of today's lesson, let's investigate the condensed model proposed by Thorn and Palmer (*Understanding Sabermetrics*, pg 87). I'll refer to this model as the Linear Weights Model.

$$BattingRuns = w_1 1\text{B} + w_2 2\text{B} + w_3 3\text{B} + w_4 \text{HR} + w_5(\text{BB} + \text{HBP}) - \underbrace{w_6(AB - H)}_{\text{Outs}}$$

How can we get weights for the Linear Weights Model?

w_1 = average run value of a single

RV = $Exp_{NEW} - Exp_{OLD}$ + Runs Scored

Exp. Runs

Runners	0	1	2
000	–	–	–
001			
⋮	–	–	
111			

How do we interpret *BattingRuns*?

The player's runs produced above the "average" player.

Let's see what this looks like for the 2018 season.

```
library(Lahman)
library(tidyverse)
library(plotly)
library(knitr)

# Calculate weights from run values using Retrosheet play-by-play data
# this code will only run if you have a 2018 retrosheet
# event-by-event data on your computer.
source("../MA388_Solutions/linear_weights.R")
weights <- linear_weights(2018) %>% pluck("weights")
weights %>% kable(digits = 3)
```

① run expectancy matrix

② calculate run value of each hit type

Event	weight
1B	0.449
2B	0.765
3B	1.097
BB.HBP	0.303
HR	1.380
Out	-0.265

1B + 2B + 3B + HR

H + BB + HBP

AB + BB + HBP + SF

How do these weights compare to those in slugging percentage (SLG)?

SLG - Double is equal to 2x a single

1x 1B + 2x2B + 3x3B + 4 HR

AB

weights

$\frac{0.765}{0.449} = 1.7$

all are weighted equally

Triples / HR weighted very heavily

OBP ⟵ ✕ ⟶ SLG

```
# 2018 Players with at least 500 at bats
vars = c("AB", "H", "X2B", "X3B", "HR", "BB",
         "HBP", "SF","RBI")
Batting %>%
  filter(yearID == 2018) %>%
  group_by(playerID) %>%
  summarise_at(vars, sum) %>%
  filter(AB >= 500) %>%
  left_join(Master %>% select(nameLast,nameFirst,playerID)) %>%
  mutate(name = paste(nameFirst, nameLast, sep = " ")) %>%
  select(name, everything(), -nameLast,-nameFirst) %>%
  mutate(X1B = H - X2B - X3B - HR,
         SLG = (X1B + 2*X2B + 3*X3B + 4*HR)/AB,
         OBP = (H + HBP + BB)/(AB + HBP + SF + BB),
         OPS = SLG + OBP,
         AVG = H/AB) -> batting.2018

#calculates Batting Runs using Thorn and Palmer's condensed Linear Weights model.
#note statistics and weights have to be in the same order
batting_runs <- function(statistics, weights){
  runs <- round(sum(statistics*weights),1)
  return(runs)
}

batting.2018 %>%
  group_by(playerID) %>%
  mutate(batting.runs = batting_runs(statistics = c(X1B, X2B, X3B, BB + HBP, HR, AB - H),
                                     weights = weights %>% pull(weight))) %>%
  arrange(-batting.runs) %>%
  group_by()-> batting.2018

batting.2018 %>%
  select(name, AB, H, HR, RBI, AVG, OPS, batting.runs) %>%
  head(10) %>%
  kable(digits = 3,
        caption = "Top 10 MLB Players (with at least 500 at bats) - Batting Runs 2018")
```

⟵ $w_1 1B + \cdots + w_6 Outs$

Table 2: Top 10 MLB Players (with at least 500 at bats) - Batting Runs 2018

"lead-off"

name	AB	H	HR	RBI	AVG	OPS	batting.runs
Mookie Betts	520	180	32	80	0.346	1.078	65.4
J. D. Martinez	569	188	43	130	0.330	1.031	58.4
Christian Yelich	574	187	36	110	0.326	1.000	52.7
Jose Ramirez	578	156	39	105	0.270	0.939	43.4
Alex Bregman	594	170	31	103	0.286	0.926	42.1
Nolan Arenado	590	175	38	110	0.297	0.935	40.1
Paul Goldschmidt	593	172	33	83	0.290	0.922	39.6
Manny Machado	632	188	37	107	0.297	0.905	35.8
Bryce Harper	550	137	34	100	0.249	0.889	35.5
Freddie Freeman	618	191	23	98	0.309	0.892	35.4

How many wins would you attribute to Mookie Betts in the 2018 season?

$\approx$ 6-7 wins

```
batting.2018 %>%
  select(name, AB, H, HR, RBI, AVG, OPS, batting.runs) %>%
  tail(10) %>%
  kable(digits = 3,
        caption = "Bottom 10 MLB Players (with at least 500 at bats) - Batting Runs 2018")
```

Table 3: Bottom 10 MLB Players (with at least 500 at bats) - Batting Runs 2018

name	AB	H	HR	RBI	AVG	OPS	batting.runs
Brian Dozier	553	119	21	72	0.215	0.696	-9.0
Nick Ahmed	516	121	16	70	0.234	0.700	-9.3
Jon Jay	527	141	3	40	0.268	0.678	-10.4
Tim Anderson	567	136	20	64	0.240	0.687	-13.7
Carlos Sanchez	600	145	8	55	0.242	0.678	-13.8
Freddy Galvis	602	149	13	67	0.248	0.680	-14.0
Amed Rosario	554	142	9	51	0.256	0.676	-14.2
Kyle Seager	583	129	22	78	0.221	0.673	-17.1
Billy Hamilton	504	119	4	29	0.236	0.626	-19.9
Dee Gordon	556	149	4	36	0.268	0.637	-21.4

Let's see how *BattingRuns* compare to traditional statistics.

```
library(gridExtra)

title.size = 12

p.rbi <- batting.2018 %>%
  ggplot(aes(label = name,
             x = batting.runs,
```

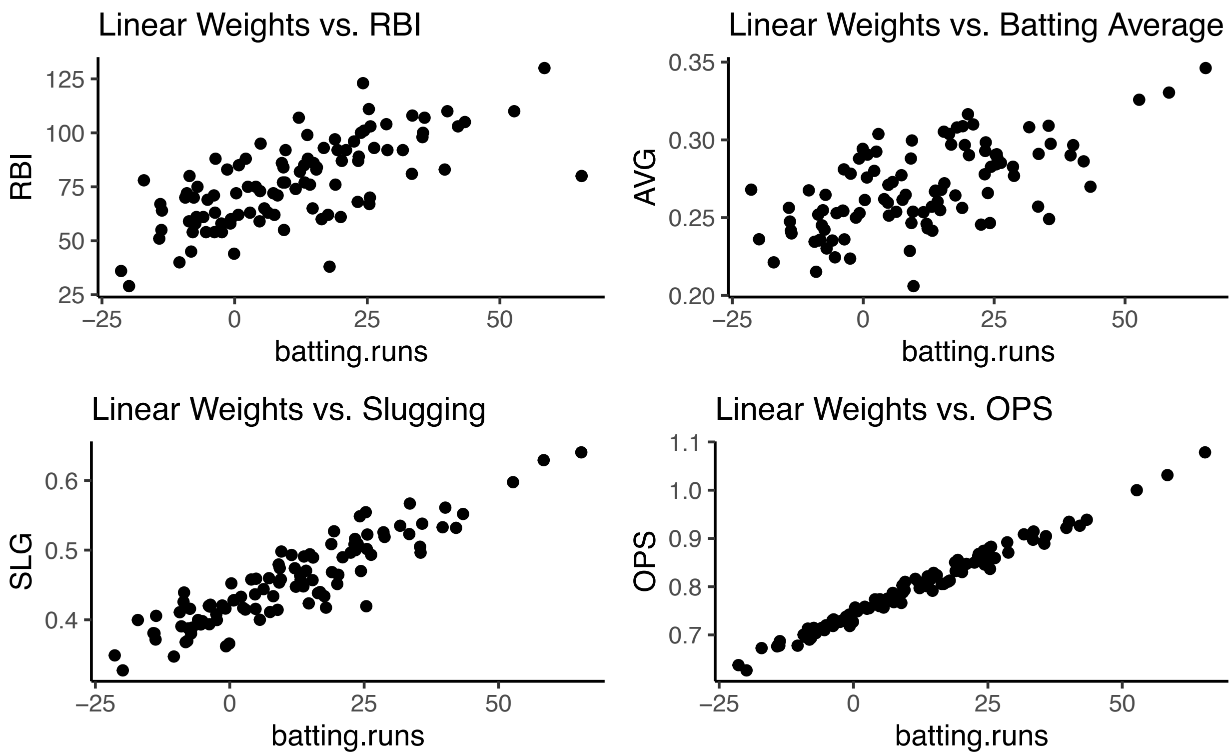
```r
             y = RBI)) +
  geom_point() +
  labs(title = "Linear Weights vs. RBI") +
  theme_classic() +
  theme(plot.title = element_text(size = title.size))

p.avg <- batting.2018 %>%
  ggplot(aes(label = name,
             x = batting.runs,
             y = AVG)) +
  geom_point() +
  labs(title = "Linear Weights vs. Batting Average") +
  theme_classic() +
  theme(plot.title = element_text(size = title.size))


p.slg <- batting.2018 %>%
  ggplot(aes(label = name,
             x = batting.runs,
             y = SLG)) +
  geom_point() +
  labs(title = "Linear Weights vs. Slugging") +
  theme_classic() +
  theme(plot.title = element_text(size = title.size))

p.ops <- batting.2018 %>%
  ggplot(aes(label = name,
             x = batting.runs,
             y = OPS)) +
  geom_point() +
  labs(title = "Linear Weights vs. OPS") + theme_classic() +
  theme(plot.title = element_text(size = title.size))

grid.arrange(p.rbi,p.avg, p.slg, p.ops, ncol = 2)
```




```
#ggplotly(p.rbi)
#ggplotly(p.avg)
#ggplotly(p.slg)
#ggplotly(p.ops)
```

Let's say you're a general manager. Which statistic would you use?

What other factors would you want to consider?

In terms of *BattingRuns*, is it more important to hit for power or on base percentage?

```r
library(viridis)
batting.2018 %>%
  mutate(OBP.scaled = scale(OBP),
         SLG.scaled = scale(SLG),
         br.scaled = scale(batting.runs)) %>%
  ggplot(aes(x = OBP.scaled, y = SLG.scaled, color = br.scaled)) +
  geom_point() +
  scale_color_viridis() +
  geom_abline(slope = 1, intercept = 0)
```