Lesson_29_Boardsheet

Kevin Cummiskey

4/8/2020

Review

Last class, we discussed Linear Weights models. Linear Weights models attempt to estimate *BattingRuns* for each player using individual performance. What are *BattingRuns*?

How do we obtain weights in the model?

What are some limitations of the Linear Weights models?

Runs Created

In the late 1970s, Bill James developed a series of statistics called *Runs Created (RC)*. These statistics have the common form:

$$RC = \frac{A \times B}{C}$$

where A quantifies how often the player/team gets on base, B quantifies how they advance on base, and C represents the opportunities. The simplest version of RC is:

$$RC_{\text{basic}} = \frac{(H + BB) \times TB}{AB + BB}$$

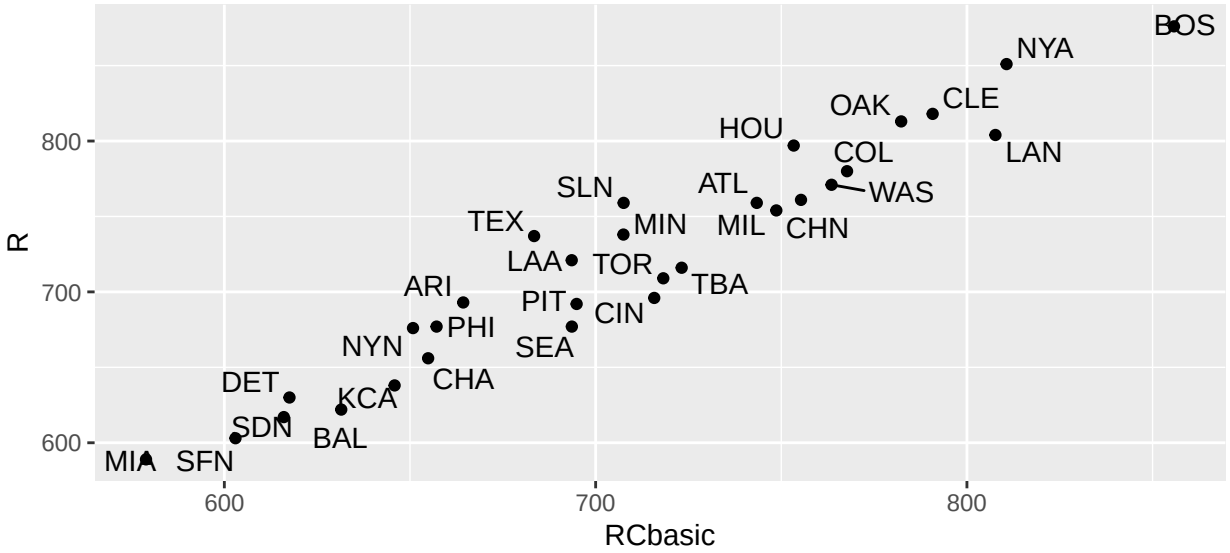
Let's see how well RC_{basic} predicts actual runs scored for teams in the 2018 season.

```
library(Lahman)
library(tidyverse)
library(ggrepel)
library(plotly)
```

```
library(knitr)

Teams %>%
  filter(yearID == 2018) %>%
  mutate(X1B = H - X2B - X3B - HR,
         TB = X1B + 2*X2B + 3*X3B + 4*HR,
         RCbasic = ((H + BB)*TB)/(AB + BB)) -> teams.2018

teams.2018 %>%
  ggplot(aes(x = RCbasic, y = R, label = teamID)) +
  geom_point() +
  geom_text_repel()
```



Is RC_{basic} a *good* model for runs? What makes it a good model?

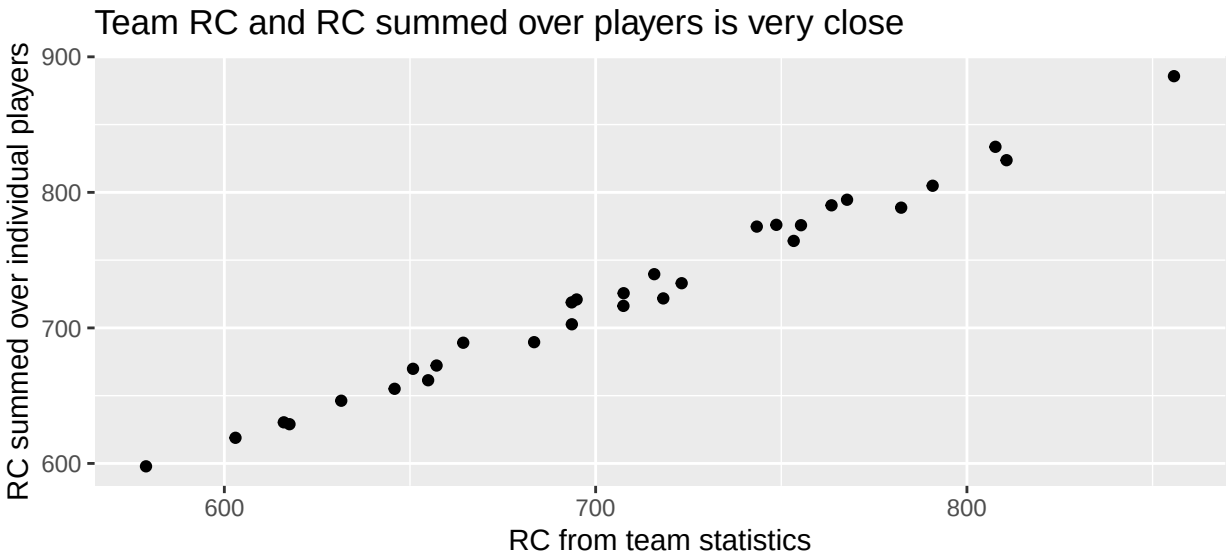
A nice feature of RC is a team's RC is very close to the sum of its individual players. Therefore, we can use RC to assign credit to players for a certain proportion of their team's runs.

```
Batting %>%
  filter(yearID == 2018) %>%
  mutate(X1B = H - X2B - X3B - HR,
         TB = X1B + 2*X2B + 3*X3B + 4*HR,
         RCbasic = ((H + BB)*TB)/(AB + BB)) %>%
  left_join(select(Master,playerID, nameLast, nameFirst)) %>%
  mutate(name = paste(nameFirst, nameLast, sep = " ")) -> batting.2018.all

batting.2018.all %>%
  group_by(teamID) %>%
  summarize(RCbasic.summed = sum(RCbasic, na.rm = TRUE)) %>%
```

```
  right_join(teams.2018) -> teams.2018

teams.2018 %>%
  ggplot(aes(x = RCbasic, y = RCbasic.summed)) +
  geom_point() +
  labs(x = "RC from team statistics",
       y = "RC summed over individual players",
       title = "Team RC and RC summed over players is very close")
```




```
batting.2018.all %>%
  left_join(select(teams.2018, teamID, RCbasic.summed)) %>%
  group_by(teamID) %>%
  mutate(RC.team.perc = RCbasic/RCbasic.summed) %>%
  arrange(-RC.team.perc) -> batting.2018.all


batting.2018.all %>%
  select(name, teamID, R, RC.team.perc) %>%
  head(10) %>%
  kable(digits = 2,
        caption = "Highest Percentage of Team's Runs Scored - 2018")
```

Table 1: Highest Percentage of Team's Runs Scored - 2018

name	teamID	R	RC.team.perc
Mike Trout	LAA	101	0.19
Paul Goldschmidt	ARI	95	0.18
Christian Yelich	MIL	118	0.18
Nick Castellanos	DET	88	0.17
Mookie Betts	BOS	129	0.16
J. D. Martinez	BOS	111	0.16
Alex Bregman	HOU	105	0.16
Nolan Arenado	COL	104	0.16

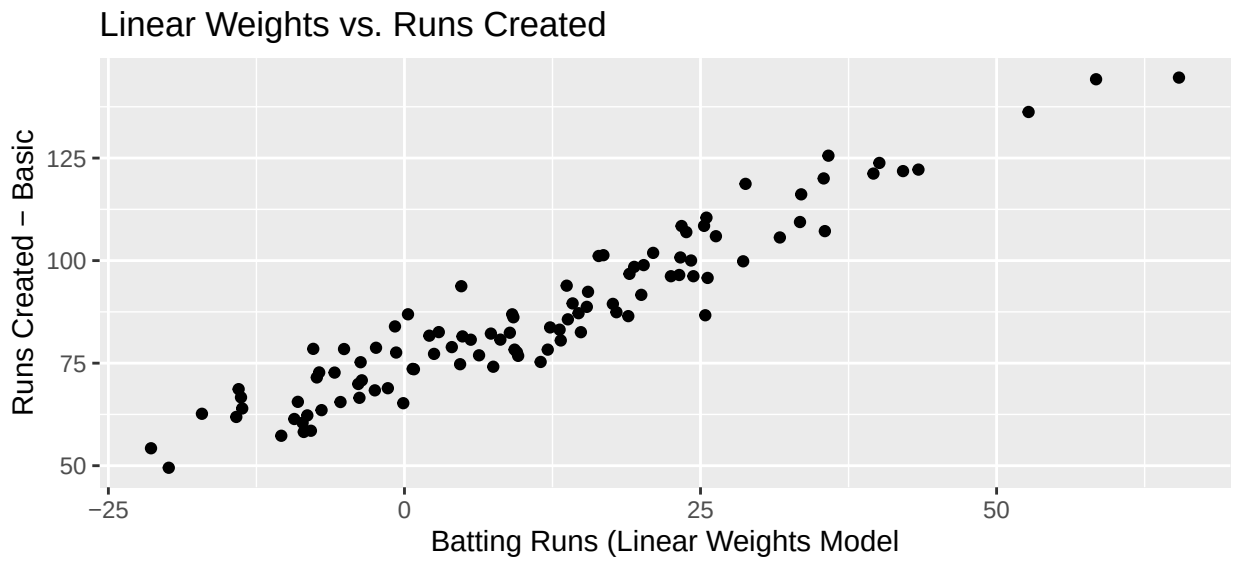
name	teamID	R	RC.team.perc
Freddie Freeman	ATL	94	0.15
Whit Merrifield	KCA	88	0.15

Let's see how *RunsCreated* compares to *BattingRuns* from the Linear Weights model.

```
read_csv(file = "data/batting_2018.csv") -> batting.2018

#add runs created to data frame from last class
batting.2018.all %>%
  group_by(playerID) %>%
  summarize(RCbasic = sum(RCbasic, na.rm = TRUE)) %>%
  right_join(batting.2018) -> batting.2018

p1 <- batting.2018 %>%
  ggplot(aes(x = batting.runs,
             y = RCbasic,
             label = name)) +
  geom_point() +
  labs(x = "Batting Runs (Linear Weights Model",
       y = "Runs Created - Basic",
       title = "Linear Weights vs. Runs Created")
p1
```




```
#ggplotly(p1)
```

Why might we prefer one model to the other?

How can we improve upon the basic version of Runs Created?