Lab 2: Metric evaluation in baseball

Michael Lopez, Skidmore College

Important note 1

First, we'll open RStudio by going to http://r.skidmore.edu/.

Important note 2

Open a new R Markdown file (File / New File / R Markdown...). You can create a basic name – Lab0, for example – and that'll set you up with a new file ready to go.

Important note 3

The base RMarkdown file is not particularly useful. Instead, click on the link below:

https://raw.githubusercontent.com/statsbylopez/FYE_18/master/HWs/HW_lab_base.Rmd

Copy and paste the text at the link above into your Markdown document.

Set the output to HTML mode, and click "Knit HTML" at the top (or Command+Shift+K on Mac). This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission. You will also print this and hand in for each homework assignment.

Lab 2: Evaluation of metrics in baseball

To get you started, enter the following in one of the code chunks.

```
library(Lahman)
library(tidyverse)
Hitter.data.16 <- Batting %>%
  filter(AB >= 400, yearID >= 2015) %>%
  mutate(so.rate = SO/(AB + BB),
         bb.rate = BB/(AB + BB),
         hr.rate = HR/(AB + BB))
```

1. What are the three newly created variables?

Understanding context

Let's create a correlation matrix that contains several hitter variables.

```
library(corrplot)
Vars.cor <- Hitter.data.16 %>%
  select(SO, BB, HR, so.rate, bb.rate, hr.rate)
cor.matrix <- cor(Vars.cor)
cor.matrix
corrplot(cor.matrix, method = "number")
```

2. Explain why it is unsurprising that three of pairs of metrics have correlations that are *way* higher than all of the other correlations. Next, explain how that may impact how we use these metrics for understanding players.

3. The code for this data set filters to players with at least 400 at bats. If we removed this filter, explain what would likely happen to the correlation betwee `SO` and `so.rate`.

4. Let's try this out. When you run the following code – which only includes an at bat threshold of 10 at bats – what happens to the correlation between strikeout rate and strikeouts?

```
Hitter.data.16.part2 <- Batting %>%
  filter(AB > 10, yearID >= 2015) %>%
  mutate(so.rate = SO/(AB + BB),
         bb.rate = BB/(AB + BB),
         hr.rate = HR/(AB + BB))

Vars.cor <- Hitter.data.16.part2 %>%
  select(SO, BB, HR, so.rate, bb.rate, hr.rate)
cor.matrix <- cor(Vars.cor)
cor.matrix
corrplot(cor.matrix, method = "number")
```

5. Your coach asks you to summarize the link between a players' strikeout rate and number of strikeouts. What would you tell the coach?

6. Explain how the following three scatter plots ties into your answer above.

```
## The following graph uses all players with at least 400 at bats
ggplot(data = Hitter.data.16, aes(x = SO, y = so.rate)) +
  geom_point()

## The following graph uses all players with at least 10 at bats
ggplot(data = Hitter.data.16.part2, aes(x = SO, y = so.rate)) +
  geom_point()

## The following graph compares at bats to strike out rate
ggplot(data = Hitter.data.16.part2, aes(x = AB, y = so.rate)) +
  geom_point()
```

The following three questions are based on the last question above

7. Metrics are said to stabilize once it is the case that extreme observations at no longer likely. At what point does `so.rate` appear to stabilize? What does that say about the strike-out rates of someone with, say, 50 at bats?

8. The Skidmore softball team played 38 games last year and the Skidmore baseball team played 37 games. If a player plays in 20 of those games and records an average of 3.5 at bats in each game, how should we feel about using that player's strikeout rate?

9. There's a slight downward slope to the chart. Why?

Links to future performance

Above, we learn a little bit about how rate metrics behave. Next, we compare their repeatability.

We are going to estimate the year-to-year correlation of each of the rate metrics using the `Hitter.data.16` data set. The first thing we do is arrange the data by player and year.

```
Hitter.data.16 <- Hitter.data.16 %>%
  arrange(playerID, yearID)
Hitter.data.16 %>% head()
```

Certain players will have two observations – that's a good thing – while others will only have one year in which they meet our criteria for inclusion. If a player only has one year of data, he won't be included in this analysis.

Next, we use the `lead` command to get a player's performance in the following year.

```
Hitter.data.16 <- Hitter.data.16 %>%
  group_by(playerID) %>%
  mutate(so.rate.next = lead(so.rate, 1),
         bb.rate.next = lead(bb.rate, 1),
         hr.rate.next = lead(hr.rate, 1))
Hitter.data.16 %>%
  head() %>%
  print.data.frame()
```

In the above set-up, we see that only two of the first six rows will be included in our year-to-year analysis.

```
Vars.cor2 <- Hitter.data.16 %>%
  ungroup() %>%
  select(so.rate, so.rate.next, bb.rate, bb.rate.next, hr.rate, hr.rate.next)
cor.matrix2 <- cor(Vars.cor2, use = "pairwise.complete.obs")
cor.matrix2
corrplot(cor.matrix2, method = "number")
```

10. Compare walk rate, strikeout rate, and home run rate in terms of predictability.
11. Is strikeout rate in one season linked to walk rate in the next season? Explain.