HW 2: Linear regression and prediction using MLB players

Stats and sports class

Fall 2019

Preliminary notes for doing HW

1. All files should be knit and compiled using R Markdown. Knit early and often! I do not recommend waiting until the end of the HW to knit.
2. All questions should be answered completely, and, wherever applicable, code should be included.
3. If you work with a partner or group, please write the names of your teammates.
4. Copying and pasting of code is a violation of the Skidmore honor code

Homework questions

Part I: Linear regression and player metrics

Return to the `Lahman` package in R, and we'll use the `Batting` data frame. Type `?Batting` for specific insight into each variable. Primarily, it's a table with 22 batting metrics. *For all questions, we'll be using the `Batting_1` data frame.

```
library(tidyverse)
library(Lahman)
Batting_1 <- Batting %>%
  filter(yearID >= 2000) %>%
  select (playerID, yearID, AB:SO) %>%
  filter(AB >= 500)
```

1. Describe the contents of `Batting_1`: that is, provide its dimensions, and what each row in the data set corresponds to.
2. When dealing with the `Teams` data set – as in our labs and prior homework – we often filtered by year. In the `Batting` data set, we are filtering by year and requiring an at-bat minimum. Why is this second step often required when working with players but not when working with teams?
3. Make a correlation matrix - both a matrix of the variables, as well as a visualization – using hits, doubles, triples, home runs, RBI, and strikeouts.
4. Make a scatter plot of runs batted in (RBI, the y-variable) and home runs (HR, the x-variable). Estimate and write the regression line using the `lm` command. Finally, interpret the slope and intercept of this line.
5. Pete Alonso – currently with the New York Mets – has hit 47 home runs and batted in 109 runs (as of Sept 15, 2019). Given his home runs, what is his residual? That is, how many more or fewer runs batted in has he hit than we'd expect given his home runs?
6. Alonso seems to have fewer runs batted in than we'd expect given his home runs. Provide a few explanations for this is the case.
7. Return to your scatter plot of RBI versus HR. Use the `annotate` command to add in a label (Alonso's name, or a symbol) with where Alonso lies. Read more about `annotate` here: https://ggplot2.tidyverse.

org/reference/annotate.html. Among players hitting Alonso's number of home runs, is his RBI total surprising?

8. Run the following code:

```
Batting_1 <- Batting_1 %>%
  mutate(K_rate = SO/(AB + BB))

ggplot(data = Batting_1, aes(x = K_rate)) +
  geom_density()

ggplot(data = Batting_1, aes(x = K_rate, colour = yearID, group = yearID)) +
  geom_density()
```

- what is `K_rate`?
- Describe the distribution of `K_rate`: e.g, what is its center, shape, and spread
- Describe how `K_rate` has changed over the last two decades. Be precise. Have the center/shape/spread changed? If so, by how much?

Part II: Predictability of player metrics

In the above example, we looked at strikeout rate – that is, the percentage of time that a player strikes out.

9. Read the article on baseball preditability here

https://blogs.fangraphs.com/basic-hitting-metric-correlation-1955-2012-2002-2012/.

What rate metrics in baseball are most repeatable? Which metrics are least repeatable?

Let's assess the repeatability of the metrics in `Batting_2`, shown below:

```
Batting_2 <- Batting_1 %>%
  mutate(HR_rate = HR/(AB + BB),
         BB_rate = BB/(AB + BB),
         RBI_rate = RBI/(AB + BB))

Batting_2 <- Batting_2 %>%
  arrange(playerID, yearID) %>%
  group_by(playerID) %>%
  mutate(HR_rate_next = lead(HR_rate),
         K_rate_next = lead(K_rate),
         BB_rate_next = lead(BB_rate),
         RBI_rate_next = lead(RBI_rate)) %>%
  ungroup() %>%
  filter(!is.na(HR_rate_next))
```

Note: The code drops the last year of a players' career – there is no future variable to look at.

10. Use (i) scatter plots and (ii) correlation coefficients to assess the year-over-year repeatability of strikeout rate, walk rate (`BB_rate`), HR rate, and RBI rate. That is, compare each metric in a players' curret year to the metric that he records in the following year. Which of these metrics is most repeatable? Which of these is least repeatable?

11. We introduced two additional ways of assessing prediction error, mean absolute error and mean squared error. Here's an example of how to code these in R.

```
Batting_2 %>%
  summarise(mae_k_rate = mean(abs(K_rate - K_rate_next)),
            mse_k_rate = mean((K_rate - K_rate_next)^2))
```

Interpret the `mae_k_rate` above. How does this number relate to the scatter plot (using `K_rate`) in Question No. 10?

12. Repeat the calculations in No. 11, only using `HR_rate` instead of `K_rate`.

13. What does the following code show?

```
ggplot(data = Batting_2, aes(x = yearID, y = HR_rate)) +
  geom_line(aes(group = playerID), colour = "grey") +
  geom_point(aes(group = playerID), colour = "grey") +
  geom_smooth()
```

14. Repeat the code in No. 13, only for K rate. Have hitters been hitting less home runs with the increase in strikeouts?