HW 4: Player prediction on MLB

Stats and sports class

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

Part I

HW Grade

Return to Homework 2 and assign yourself a grade:

- 1-3 out of 5 points: Most questions attempted, minimal effort
- 4 of 5 points: All questions attempted, complete effort, graded questions incorrect
- 4.5 of 5 points: All questions attempted, complete effort, graded questions partially correct
- 5 of 5 points: All questions attempted, graded questions perfect

Solutions to HW 3 posted on Github

Homework questions

Part II: Multiple regression and player metrics

Run the following code to create data for this week's HW.

```r
library(tidyverse)
library(Lahman)
Batting_1 <- Batting %>%
  filter(yearID >= 1995, yearID <= 2015, AB >= 550) %>%
  mutate(K_rate = SO/(AB + BB),
         BB_rate = BB/(AB + BB),
         BA = H/AB,
         HR_rate = HR/(AB + BB),
         X1B = H - X2B - X3B - HR,
         TB = X1B + 2*X2B + 3*X3B + 4*HR,
         RC = (H + BB)*TB/(AB + BB)) %>%
  arrange(playerID, yearID) %>%
  group_by(playerID) %>%
  mutate(BB_rate_next = lead(BB_rate)) %>%
  filter(!is.na(BB_rate_next)) %>%
  ungroup()
```

```r
head(People)

Batting_2 <- Batting_1 %>%
  left_join(People) %>%
  select(playerID, birthYear, yearID, K_rate, BB_rate, HR_rate, RC, weight,
         height, bats, nameFirst, nameLast, BB_rate_next)

head(Batting_2)
```

Question 1

Read this awesome cheat-sheet about how to join data frames in R. Link: https://stat545.com/join-cheatsheet.html

Describe the difference between `left_join`, `inner_join`, and `right_join`. Next, why was `left_join` used in the code above? What variables were added to the `Batting` data frame?

Question 2

Three plots are shown below. Each one is a version of a *spaghetti* plot, called as such because of what it often appears.

```r
## Plot 1
ggplot(data = Batting_2, aes(yearID, BB_rate, group = playerID)) +
  geom_line(colour = "grey") +
  geom_point(colour = "grey")


## Plot 2
ggplot(data = Batting_2) +
  geom_line(colour = "grey", aes(yearID, BB_rate, group = playerID)) +
  geom_point(colour = "grey", aes(yearID, BB_rate, group = playerID)) +
  geom_smooth(data = Batting_1, aes(yearID, BB_rate))
```

- What does each line correspond to in each plot?
- What does the second plot highlight?

Question 2

Make one spaghetti plot for `K_rate` and `HR_rate`, and describe the trends over time for each variable.

Question 3

Identify if there are any interesting links between player characteristics such as height and weight and their on-field performances. No more than 2 plots are needed. Answers may vary.

Question 4

One critical question for teams is the impact of age on player performance. Without any analysis, describe how you would anticipate age impacting `RC` (runs created) in our baseball data set.

Question 5

```
Batting_2 <- Batting_2 %>%
  mutate(player_age = yearID - birthYear,
         player_age_sq = player_age^2)
```

The code above creates a new variable, `player_age`, the identifies the age of each player in each season. How is age linked to RC in the `Batting_2` data set? Is this surprising? Provide the primary reason that our approach for estimating the link between age and runs created is flawed.

Question 6

Fit two models to assess the link between age and walk rate.

Model 1 should assume a linear association.

Model 2 should assume a quadratic association, using `player_age_sq` in addition to `player_age`.

Which model fits best? Provide *three* ways of supporting your answer.

Question 7

Use code from our most recent lab to calculate the mean absolute error and the mean squared error for Model 1 and Model 2 above. Interpret the Mean Absolute Error for Model 2.