HW 4: Player prediction on MLB

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: Multiple regression and player metrics

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

```
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()


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.

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

## Plot 3
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 third plot highlight – albeit with some fancier code – that the first two plots miss?

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.

Part II: Open ended

A coach is hoping to predict `BB_rate` – the percentage of at bats a player has in each season that end in a walk (termed a base on balls).

Instead of our typical approach – which fits a model within a data frame, and estimates projections within that same data frame, we are going to take a different approach.

Specifically, we are going to create two data frames:

```
set.seed(0)
Batting_2 <- Batting_2 %>%
  mutate(random_seed = rnorm(nrow(Batting_2)))

training_data <- Batting_2 %>%
  filter(random_seed < .5)

test_data <- Batting_2 %>%
  filter(random_seed > .5)

dim(training_data)
dim(test_data)
```

The `training_data` will be the data set where we'll evaluate several models – the `test_data` is where those models will be compared to one another based on prediction accuracy.

Consider the following example:

```
fit_1 <- lm(BB_rate_next ~ BB_rate, data = training_data)
fit_2 <- lm(BB_rate_next ~ BB_rate + HR_rate, data = training_data)

test_data <- test_data %>%
  mutate(BB_rate_p1 = predict(fit_1, test_data),
         BB_rate_p2 = predict(fit_2, test_data))

head(test_data) %>% select(BB_rate, BB_rate_next, BB_rate_p1, BB_rate_p2)
```

Two preditions are generated – `BB_rate_p1` and `BB_rate_p2`, but instead of being used within the training data, they are being compared in data that the model has not yet seen.

Question 7

Compare the mean absolute error for the predictions from `fit_1` and `fit_2` in the test data (in other words, compare the MAE between each prediction with `BB_rate_next`) Which one is more accurate?

Question 8

Develop a series of models designed to predict `BB_rate_next`. Your goal should be to predict this variable **in the test set**, as opposed to in the training set. How low can your out of sample MAE drop to?

Question 9

A naive model would assume that every players' walk rate is identical – in the training data, this is around 0.0924. The following code provides MAE using this naive rate.

```
naive_rate <- training_data %>% summarise(ave_bb = mean(BB_rate_next))
naive_rate
test_data %>%
  summarise(ave_error_naive = mean(abs(BB_rate_next - .0924)))
```

- Interpret the `ave_error_naive` above.
- How much improvement was your model in Question 8 able to improve over this naive approach?

Question 10

In 3 to 4 non-technical sentences, describe your work in Part II to a coach. What would you tell the coach about how you can predict walk rate? And why should that matter to the coach? Again, non-technical words only.

Question 11

There's immense value in using different splits of the data (a training and a test data set) to evaluate predictions. What is one benefit? Several answers here will work.