HW 4: Player prediction on MLB

Stats and sports class

Fall 2019

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()

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)


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

Question 5

Provide the primary reason that our approach for estimating the link between age and runs created is flawed.

Answer: We're only observing players who actually got to play – and take 500 at bats or more – which means that the players that weren't good enough weren't in our sample. It's likely that several of the players we are dropping are the yonger and older players, making it appear like there's no strong impact of age.

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.

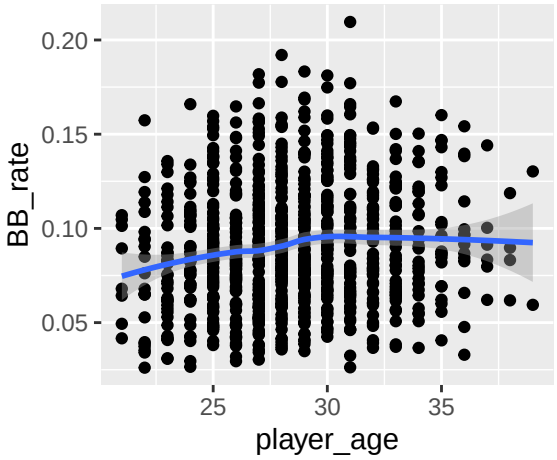
```
model_1 <- lm(BB_rate ~ player_age, data = Batting_2)
model_2 <- lm(BB_rate ~ player_age + player_age_sq, data = Batting_2)
AIC(model_1)
```

```
## [1] -3805.025
```

```
AIC(model_2)
```

```
## [1] -3807.524
```

```
ggplot(Batting_2, aes(player_age, BB_rate)) + geom_point() +
  geom_smooth()
```




```
library(broom)
tidy(model_2)
```

```
## # A tibble: 3 x 5
##   term           estimate std.error statistic p.value
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   -0.0615   0.0566        -1.09  0.277
## 2 player_age     0.00949  0.00393        2.41  0.0160
## 3 player_age_sq -0.000144 0.0000677     -2.12  0.0342
```

Answers (3 of the 4 for full credit):

1. The AIC is lower for Model 2, insinuating it's a better fit
2. In the scatter plot, there appears to be a small, negative u-shaped link between age and walk rate.
3. In `model_2`, the coefficient on the `player_age_sq` term is significant.
4. Given what we know about how age likely impacts player performance, it's safe to say that walk rate will eventially drop.

Part II: Open ended

```
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)
```

```
## [1] 648  16
```

```
dim(test_data)
```

```
## [1] 291  16
```

```
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)
```

```
## # A tibble: 6 x 4
##   BB_rate BB_rate_next BB_rate_p1 BB_rate_p2
##     <dbl>        <dbl>      <dbl>      <dbl>
## 1  0.148        0.153      0.135      0.133
## 2  0.154        0.159      0.140      0.135
## 3  0.159        0.181      0.143      0.138
## 4  0.0840       0.0598     0.0873     0.0931
## 5  0.105        0.119      0.103      0.100
## 6  0.125        0.145      0.118      0.114
```

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?

Answers

```
test_data %>%
  summarise(mae_fit_1 = mean(abs(BB_rate_p1 - BB_rate_next)),
            mae_fit_2 = mean(abs(BB_rate_p2 - BB_rate_next)))
```

```
## # A tibble: 1 x 2
##   mae_fit_1 mae_fit_2
##       <dbl>     <dbl>
## 1    0.0176    0.0173
```

The out of sample error on `fit_2` is lower – 1.73 percent compared to 1.76 percent.

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.

Answers will vary