HW 4: Player prediction on MLB

Stats and sports class

Fall 2019

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.

```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()


head(People)
```

```
##    playerID birthYear birthMonth birthDay birthCountry birthState  birthCity
## 1 aardsda01      1981         12       27          USA         CO     Denver
## 2 aaronha01      1934          2        5          USA         AL     Mobile
## 3 aaronto01      1939          8        5          USA         AL     Mobile
## 4  aasedo01      1954          9        8          USA         CA     Orange
## 5  abadan01      1972          8       25          USA         FL Palm Beach
## 6  abadfe01      1985         12       17         D.R.  La Romana  La Romana
```

```
##   deathYear deathMonth deathDay deathCountry deathState deathCity nameFirst
## 1        NA         NA       NA         <NA>       <NA>      <NA>     David
## 2        NA         NA       NA         <NA>       <NA>      <NA>      Hank
## 3      1984          8       16          USA         GA   Atlanta    Tommie
## 4        NA         NA       NA         <NA>       <NA>      <NA>       Don
## 5        NA         NA       NA         <NA>       <NA>      <NA>      Andy
## 6        NA         NA       NA         <NA>       <NA>      <NA>  Fernando
##   nameLast        nameGiven weight height bats throws      debut  finalGame
## 1  Aardsma      David Allan    215     75    R      R 2004-04-06 2015-08-23
## 2    Aaron      Henry Louis    180     72    R      R 1954-04-13 1976-10-03
## 3    Aaron       Tommie Lee    190     75    R      R 1962-04-10 1971-09-26
## 4     Aase   Donald William    190     75    R      R 1977-07-26 1990-10-03
## 5     Abad    Fausto Andres    184     73    L      L 2001-09-10 2006-04-13
## 6     Abad Fernando Antonio    220     73    L      L 2010-07-28 2019-09-28
##    retroID   bbrefID  deathDate  birthDate
## 1 aardd001 aardsda01       <NA> 1981-12-27
## 2 aaroh101 aaronha01       <NA> 1934-02-05
## 3 aarot101 aaronto01 1984-08-16 1939-08-05
## 4 aased001  aasedo01       <NA> 1954-09-08
## 5 abada001  abadan01       <NA> 1972-08-25
## 6 abadf001  abadfe01       <NA> 1985-12-17
```

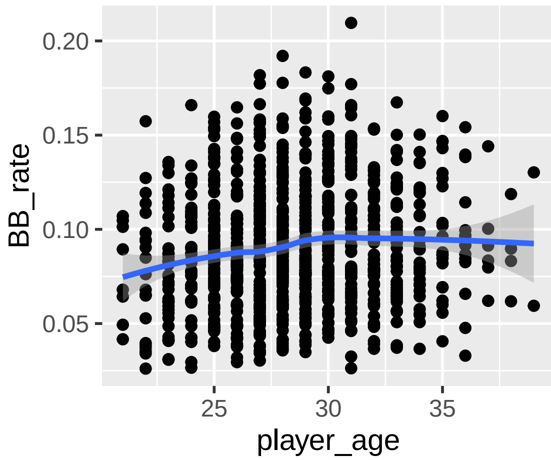
```r
Batting_2 <- Batting_1 %>%
  left_join(People) %>%
  select(playerID, birthYear, yearID, K_rate, BB_rate, HR_rate, RC, weight,
         height, bats, nameFirst, nameLast, BB_rate_next)

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

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
```

```r
AIC(model_2)
```

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

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




```r
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.