AIC, MSE, MAE, and non-linearity

Michael Lopez, Skidmore College

Overview

In this lab, we'll try and build models to predict player performance in the following season. We're going to start by using the `Batting` data.

```
library(Lahman)
library(tidyverse)
```

```
Batting_1 <- Batting %>%
  filter(yearID >= 1970, AB >= 500) %>%
  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))

Batting_1 <- Batting_1 %>%
  arrange(playerID, yearID) %>%
  group_by(playerID) %>%
  mutate(RC_next = lead(RC)) %>%
  filter(!is.na(RC_next)) %>%
  ungroup()

head(Batting_1)
```

Categorical variables

The following code creates categories for hitters based on the number of stolen bases they record in a season.

```
Batting_1 <- Batting_1 %>%
  mutate(SB_category = case_when(SB > 25 ~ "Fast",
                                 SB > 5 ~ "Moderate",
                                 SB <= 5 ~ "Slow"))

Batting_1 %>% count(SB_category)
```

The `count()` command creates a table with the frequencies of batters in each category.

A coach fits the following regression model

```
fit_run <- lm(RC ~ BB_rate + HR_rate + K_rate + SB_category, data = Batting_1)
summary(fit_run)
```

1. Interpret the coefficient on walk rate. *Note*: it's difficult to interpret, so instead of considering a 1 unit increase, consider a 0.01 (1 percent) unit increase.

2. Interpret the coefficients `SB_categoryModerate` and `SB_categorySlow`.
3. Start with the following model. Consider adding or subtracting variables until you can no longer lower the AIC

```
Batting_1 %>% head()
fit_run <- lm(RC_next ~ RC + BB_rate + HR_rate + K_rate + SB_category, data = Batting_1)
summary(fit_run)
```

4. The following code will estimate MAE and MSE for your model above. Interpret each of these numbers.

```
Batting_1 %>%
  ungroup() %>%
  mutate(rc_predict = predict(fit_run, Batting_1)) %>%
  summarise(mse = mean((rc_predict - RC_next)^2),
            mae = mean(abs(rc_predict - RC_next)))
```

Linear models with non-linear terms

The association between home run rate (`HR_rate`) and `RC_next` is kind of funky.

```
ggplot(Batting_1, aes(HR_rate, RC_next)) + geom_point()
ggplot(Batting_1, aes(HR_rate, RC_next)) + geom_point() + geom_smooth()
ggplot(Batting_1, aes(HR_rate, RC_next)) + geom_smooth()
```

One way to account for the curved nature of the association is to include a quadratic term in the regression model.

```
fit_1 <- lm(RC_next ~ HR_rate, data = Batting_1)

Batting_1 <- Batting_1 %>%
  mutate(HR_rate_sq = HR_rate^2)

fit_2 <- lm(RC_next ~ HR_rate + HR_rate_sq, data = Batting_1)
library(broom)
tidy(fit_2)
AIC(fit_1)
AIC(fit_2)
```

5. Does it make sense to include the quadratic term in the model?
6. Why is the coefficient on the quadratic term negative?
7. Can the coefficient on `HR_rate` be interpreted as we usually do it?
8. Scatter plots help justify the quadratic term, but we should check the other two assumptions for model fitting. Do so here, and compare the two fits above