Multiple regression and R-squared

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.

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

```r
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),
         lgID_next = lead(lgID)) %>%
  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.

```r
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

```r
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. Consider the context in baseball – what do you think is responsible for the coefficients you are observing in Question 2.

Comparing multiple regression models.

Ultimately, baseball coaches are tasked with predicting performance in the following season. Our goal today is to predict runs created in the next year `RC_next`.

4. Use several scatter plots to estimate how a few variables are linked to `RC_next`.
5. Create several multivariate regression models, using any set of input to guide you. your outcome must be `RC_next`.
6. Evaluate the models in No. 4 using the AIC criterion.
7. For the model with the lowest AIC in No. 5, generate a set of predictions for each player. Call these predictions `RC_next_predict`.
8. The first row in `Batting_1` is Bobby Abreau, and it corresponds to the 1999 season. In the 2000 season, Abreau's `RC_next = 133.074`. What is your prediction (`RC_next_predict`) for Abreau in that season?

```
Batting_1 %>% slice(1) %>% print.data.frame()
```

9. See our lecture notes – calculate the mean absolute error and mean squared error for the entire set of `RC_next_predict`. You should only be doing this for one model.
10. Interpret the mean absolute error in No. 9. What does it say about your runs created predictions?
11. Compare the distribution of your entire set of `RC_next_predict` values to the observed `RC_next` values using a scatter plot. What does this say about the appropriateness of your model?

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)
```

12. Does it make sense to include the quadratic term in the model?
13. Why is the coefficient on the quadratic term negative?
14. Can the coefficient on `HR_rate` be interpreted as we usually do it?