HW 3: Multivariate regression in 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

Return to the `Lahman` package in R, and we'll use the `Teams` data frame. Below, we create a variable for the number of singles each team had in a season.

```
library(tidyverse)
library(Lahman)
Teams_1 <- Teams %>%
  filter(yearID >= 2000) %>%
  mutate(X1B = H - X2B - X3B - HR)
```

Question 1

Let's use the `Teams` data set (recall: to load this data set from the Lahman package, run the command `data(Teams)`). Using every season since 2000, fit a multiple regression model of runs (`R`) as a function of singles, doubles, triples, home runs, and walks. Showing the code output is sufficient for this question.

```
#Your code goes here
```

Question 2

Refer to the fit in question 1. Identify the y-intercept, as well as the slopes for singles, doubles, triples, home runs and walks (do not interpret).

Question 3

Refer to the fit in question 1. Interpret the slope coefficient estimate for triples.

Question 4

Use the fit in question 1 to generate a set of predicted runs scored for each team in your data set.

What is the correlation between your predicted runs and the number of actual runs?

Question 5

Identify the following:

i) The number of runs scored by Anaheim in 2000 (`teamID == "ANA"`)

ii) The predicted number of runs scored by Anaheim in 2000, using your model in question 1.

Question 6

Using `mutate()`, create a new variable for the residual between the observed number of runs for each team and what your model predicted.

Next, answer:

i) Which team-season corresponds to the highest residual?

ii) Plot the residuals versus `yearID`: Is there any pattern? Would `yearID` be an appropriate term to add to the model?

Question 7

Using the output from Question 1, discuss the relative importance of each type of productive at bat (singles, doubles, triples, home runs, walks) with respect to run generation. Does anything surprise you?

Question 8

Pick another variable in the `Teams` data set, and add it to your regression model. Interpret it's slope. Also, does this new variable appear to be significantly associated with runs scored, given the other variables in the model?

Part II: Model assessment

Several models are proposed.

```
fit_1 <- lm(R ~ X1B + X2B + X3B + HR, data = Teams_1)
fit_2 <- lm(R ~ X1B + X2B + X3B + HR + BB, data = Teams_1)
fit_3 <- lm(R ~ X1B + X2B + X3B + HR + BB + SO, data = Teams_1)
fit_4 <- lm(R ~ X1B + X2B + X3B + HR + BB + SO + CS, data = Teams_1)
fit_5 <- lm(R ~ X1B + X2B + X3B + HR + BB + SO + CS + lgID, data = Teams_1)
fit_6 <- lm(R ~ X1B + X2B + X3B + HR + BB + SO + CS + lgID + SB, data = Teams_1)

options(scipen=999)
```

Note: The `options(scipen = 999)` command disables R's scientific notation.

Question 9

Using the AIC criteria, which of the six models would you recommend for measuring runs scored on a team-wide level? From a baseball perspective, what does your choice suggest about certain measurements as far as their link to runs scored?

Question 10

One of the coefficients in `fit_5` and `fit._6` is `lgID`. Generate a table of the `lgID` in your data set. What does this variable refer to?

Question 11

Using the code below, the coefficient for `league = "NL"` is negative. Interpret this coefficient. What about baseball's rules make it important to consider which league each team played in? Note: you can google the differences between the American League and the National League to guide you.

```
library(broom)
tidy(fit_5)
```