Problem Set 1: An Introduction to R
DSCI 304

Instructions: Answer the following questions, using R (where necessary) to assist you. Type your answers to the questions in a Word document. In addition, create a text file (something like "StudentNameAssignment1.r") that contains all the commands that you used to answer the questions. Submit both the word and the text script file with R commands (or a markdown file containing both your commands and text responses) to Canvas.

Examining Batters in Major League Baseball

Use the read.csv command to open the MLB Batting Data.csv dataset. This is a dataset containing the batting records and statistics from 729 major league hitters between 2015 and 2019. Use this dataset to perform the following tasks and answer the associated questions. Specific variable names you will need are included are included in parenthesis.

1) There is a longstanding saying in baseball that having a career batting average (batting_avg) of 0.300 (that is, earning a hit on at least 30% of your at bats) will earn you a place in the Baseball Hall of Fame. If that saying is correct, how many of the players in this dataset can expect to make the hall of fame?[1]

2) When up to bat, players are sometimes hit by a pitch. Using the hit by pitch variable (b_hit_by_pitch), create a new variable that is a 0 if the player was never hit by a pitch in a given season, and a 1 if the player was hit at least once. Report a table of your new variable. How many players were not hit at least once a season?

3) It is commonly thought that the harder you hit a baseball, the more likely you are to score a homerun. Assess the validity of that claim: Run a linear regression examining the effect of a players average exit velocity (exit_velocity_avg) on the number of homeruns they hit (b_home_run). What do you find? Is the conventional wisdom correct?

[1] To make the analysis of this data easier, you are free to assume that each player-year is a different player (ie that Albert Pujols in 2015 is a completely different person than Albert Pujols in 2016, and each could potentially make the hall of fame). You are also free to make similar assumptions for all of the questions in this problem set.

4) Age can be an important limitation for a hitter; it is often thought that older hitters perform at lower levels than their younger teammates. Assess that claim: Run a linear regression examining the effect of a player's age (player_age) on their batting average (batting_avg). Discuss what you find. Does age seem to play an important factor?

   a. It may be the case that age really only matters after important milestones. Using the player_age variable, create a new variable that is a 0 if the player is less than 30 years old, and a 1 if the player is at least 30 years of age. Repreform your previous analysis using your new variable instead of the player_age variable. What do you find? Does this variable change the results or your interpretation of the importance of age?

5) As an alternative to the traditional batting average stats, Major League Baseball has a separate stat known as the Slugging Percentage, which is a weighted average that gives more weight to extra base hits (hits where the batter reaches more than first base). Unfortunately, this statistic is not included in our data set. Using the formula below, and the variables for single base hits (b_single), double base hits (b_double), triple base hits (b_triple), homeruns (b_home_run), and at bats (b_ab), calculate each players Slugging Percentage. What is the average Sluggging Percentage in the dataset? How many players have slugging percentages above .500?

$$Slugging\ Percentage = \frac{(Single\ Base\ Hits)+(2*Double\ Base\ Hits)+(3*Triple\ Base\ Hits)+(4*Home\ Runs)}{At\ Bats}$$

6) Next, subset the full dataset into a new data frame that only includes observations from the 2019 season.

   a. How many players (ie rows) from 2019 are included in the data?
   b. What is the average number of plate appearances (b_total_pa) taken per player in 2019?
   c. What was the most number of homeruns (b_home_run) hit by a player in 2019? What player hit those homeruns?