Free Assignment 4 - Regression and multiple regression in Baseball

The intersection of statistics and sports has changed how we see sports in the last 40 years. This probably first happened in baseball with statistics articles published by a series of fans in the journal SABR (leading to the term sabrmetrics) and was captured in the book and movie *Moneyball*.

One idea that comes up a lot in statistics research in sports is how to identify the most important skills in a game. In baseball, the objective is to score runs and keep the other team from scoring runs. There are many ways to score runs, however, and some ways may be better than others. For instance, some batters are good at getting hits, others may be good at getting on base and avoiding getting out. Some get out a lot but hit many home runs. Knowing which of these skills is most important can help teams pick the right players.

In this assignment, we are going to use multiple regression to determine which skills are most likely to lead to team success in offense. The objective of a team is to score runs.

When doing this assignment, make sure your R notebook contains all your analyses, and all the code to complete the required **bolded** steps below, as well as the answers for the questions.

Step 1: Create a folder, a new R notebook, and download the 'FA4_baseball.csv' data into the folder. Load the data into a dataframe with the name `b` for baseball. Also, load the tidyverse package.

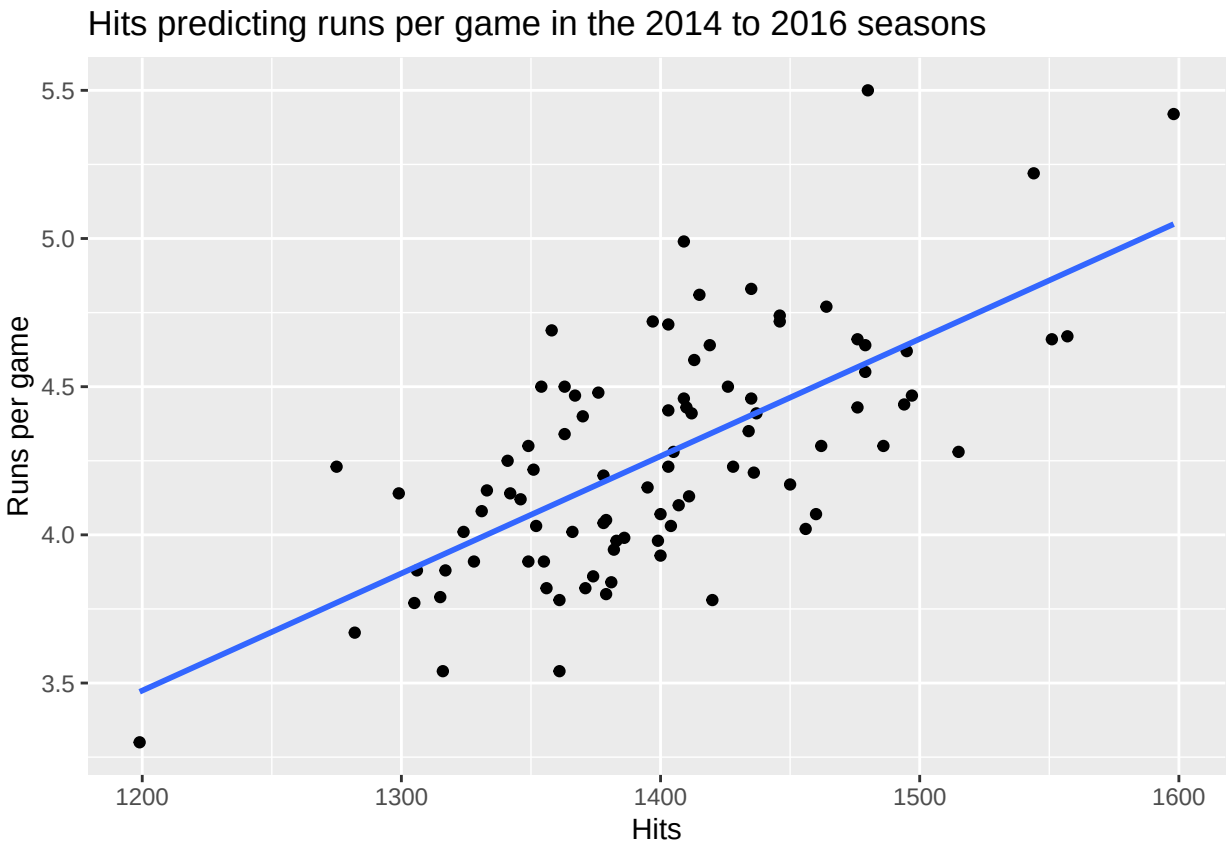
When you load the package, check to make sure it inputted correctly. There should be 90 observations of 31 variables. The observations are for the 2014-2016 seasons and here is a guide to some of the column names. You can find out what all the column names are here.

- Tm: team name
- League: League (AL for American League, NL for National League)
- Year: year data were collected
- BatAge: The average age of the team's batters
- RpG: Runs per game (this will be our main outcome variable)
- R: total runs in a season
- H: Total hits in a season
- 2B: Total number of doubles
- 3B: total number of triples
- HR: total number of Home Runs
- SB: total number of stolen bases
- BB: total number of walks
- SO: total number of strikeouts
- BA: Batting average (number of hits divided by number of at bats)
- OBP: On-base percentage (percent of time a person reaches base)

Step 2: First, we want to visualize the relationship between some of our predictor variables and runs per game. Create scatterplots using ggplot which visualize the following relationships. Also add a regression line to each of the plots. Make sure your axes and plot have ap propriate titles:

A. Hits predicting Runs per game B. HR predicting Runs per game C. SB predicting Runs per game D. Batting average predicting Runs per game.

Your first plot should look something like this:



Step 3: In your annotation, answer the following question: When looking at the plots, do you see any outliers? What do you think the plots tell you?

Now we are going to do a concept called *stepwise regression*. The idea of this is to start out with one variable and add other variables and watch to see how the relationship of each predictor changes as we add more predictors.

Step 4: Create three regressions, using the following predictors:

A. Hits predicting Runs per game B. Hits and Walks predicting Runs per game C. Hits, walks and HR predicting Runs per game

Step 5: When looking at each of these regressions, what do you think the story is? Which do you think is most important in predicting runs per game?

Now we want to look at each league separately.

Step 6: Using what you learned in the tidyverse lab, create two data frames, one data frame containing only NL teams and one data frame containing only AL teams.

Step 7: Repeat the work you did for Step 3 for each of these new data frames and see if there are any differences between the AL and NL in the regressions. Are there any differences between the AL and NL in the regressions? What differences might there be?

Step 8: Think of another variable that you think might predicts runs per game (other than hits). Does it predict runs per game? Does it still predict runs per game when controlling for Hits?