NURSG 741 Project

Lacey Gleason

2/14/2018

Milestone 1

Basic Information 

Project Title: Predictors of Major League Baseball Player Injury in 2017
Author Name: Lacey Gleason
Email Address: lpgleas@emory.edu

Overview and Motivation 

I am interested in undertaking this project because I am a big baseball fan. In addition to causing diminished health status and loss Since baseball player salaries are guaranteed, injuries can be especially costly for teams. As there are many factors involved, I am interested in better understanding what player characteristics are associated with placement on the disabled list in an upcoming season. I also picked this project because it will require me to develop a number of skills including web scraping and data wrangling from a number of large databases with layered data.

Project Objectives 

The focal question that I am trying to answer is what Major League Baseball player characteristics were associated with placement on the disabled list during the 2017 season. I would like to learn how to conduct web scraping by gathering transaction data from the MLB website, which allows this practice. I would also like to improve my data wrangling skills in R by combining data from multiple large data sources, which will require pulling data into R, assessing missing information, collapsing data to create summary measures for certain player characteristics, merging data, and creating a model for probability of placement on the disabled list and number of days on the disabled list among those who were placed on the disabled list.

Player Characteristics of Interest

- **All Players**
  - Games played in 2016
  - Body Mass Index (calculated from height and weight) in 2017
  - Age at beginning of 2017 season

  - Throws right
  - Bats right
  - Number of different positions played in 2016
  - Main position in 2016
  - Home park in has artificial turf in 2016
  - Home park in has artificial turf in 2017
  - New team in 2017
  - 2017 is last year prior to free agency
  - American League
  - Number of days on disabled list in 2016
- **For Pitchers Only**
  - Pitch count in 2016
  - Pitch breakdown in 2016
  - Pitch type breakdown in 2016
  - Average fastball velocity in 2016
  - Ever had arm or shoulder surgery

Data Source 

The data for this project will come from two main sources, the record of transactions on Major League Baseball's website http://mlb.mlb.com/mlb/transactions and baseball statistics compiled in the Baseball Reference Team Rosters http://Baseball-Reference.com. Where data are missing, supplemental data on player characteristics may be gathered from the player section of Major League Baseball's website and supplemental raw game data may be gathered from Retrosheet http://www.retrosheet.org/boxesetc/index.html#Players.

Data Source	Data Elements
MLB.com	Home stadium, disabled list dates, type of injury, movement to new team, pitch location breakdown, pitch type breakdown, pitch velocity
Baseball Reference	Height, weight, date of birth, throwing arm, batting side, games played, games played at each position, contract year information, pitch count
Retrosheet	Missing data as needed - Retrosheet contains very detailed play-by-play information

Data Wrangling 

Yes, I anticipate that there will be extensive data cleaning, reshaping, and extraction tasks. I will need to scrape the transaction data from MLB.com using the rvest package. It will take some time to familiarize myself with the package documentation and the HTML script on the website in order to isolate the elements needed for my project. Once those data elements have been brought into R, the data will need to be cleaned and reshaped so that the data are collapsed by player rather than listed by date of transaction. Player

characteristics of interest will then be pulled in from the other databases and matched to the transaction data by player identification numbers.

Analysis 

The analytic sample will include all players who played in the MLB in both 2016 and 2017. I will report descriptive statistics on player characteristics. A two-part model will be utilized in this analysis. The first part of the model will be a logistic regression model predicting whether or a player was put on the disabled list in 2017. I will then create a two-part model to predict likelihood of a player being placed on the disabled list in 2017. The first part of the model will be a logistic regression with the outcome of whether or not the player was placed on the disabled list in 2017. The second part of the model will not include the observations that were zeros for disabled list placement in 2017.The second model will assess the discrete count variable of the days on the disabled list in 2017 among those who had any days on the disabled list. The appropriate distribution (i.e., family) and link to choose for the second part of the model will be determined by running a Modified Park test. Overall, this two-part analysis will provide condensed marginal effects and takes into account the original distribution which has many zeros. A sub-analysis will be conducted of pitchers including the pitching specific characteristics above. All analyses will be conducted in R Studio Version 1.1.419 – © 2009-2018 RStudio, Inc.

Schedule of Weekly Tasks/Goals 

Week	Task
February 12	Complete Milestone 1, Schedule meeting to discuss proposal
February 19	Complete web scraping of MLB transactions data
February 26	Clean transaction data, aggregate data by player, write formulas to calculate days on disabled list, number of times on disabled list, body part associated with disabled list transaction
March 5	Pull data from Baseball Reference on player characteristics of interest and add to disabled list data matching by player ID
March 12	Write code for descriptive statistics, set up model, submit working prototype
March 19	Create Table 1 for descriptive statistics, create tables/graphs for model results
March 26	Write introduction/background section of manuscript
April 2	Write methods/results section of manuscript
April 9	Write discussion/conclusions section of manuscript
April 16	Submit manuscript
April 23	Submit presentation

Link 

The link for this file is located at https://github.com/lpgleason/Project.git.