Nick Chapman

MAT 8406

Dr. Yimin Zhang

03/07/2020

Project Proposal

1. Empirical Question

Major League Baseball is notably among the most data-rich institutions in sports, if not among all industries. Even prior to the rise of automated data-gathering, fans of the sport took it upon themselves to operationalize the nuanced details of the game, and with the advent of the personal computer, it became possible for many to contribute. Bill James was essentially the founder of a grass roots movement to collect baseball data not just for current games, but data from film of prior seasons. From this manual data entry, there has evolved many different, smarter ways of collecting data. Famously, PITCHf/x, “a camera-based, motion-tracking system […] [that] measure[s] if pitches were in the strike zone (Sawchik 60), became a new way to measure novel metrics such as pitch type and ball spin rate. With all 30 MLB parks installing the system in 2008, over a decade of data is now publicly available that drills down to pitch level. While this data is used to evaluate various performance KPIs for game strategy such as a pitcher’s best pitches, or a batter’s favorite pitch type, I believe no application of this data is more ethical and innovative than in evaluating pitcher injury. Using this PITCHf/x data, and detail on player’s injuries, I want to investigate if there is any relationship between pitch type breakdown, average velocities, age, innings pitched, pitches thrown, etc., and pitching injuries. Specifically, I want to investigate, is there any relationship between these variables and pitchers who eventually require Tommy John Surgery, named for the first pitcher who underwent the operation.

1. Importance

Promising young athletes often make sacrifices with the dedication to their craft. They provide entertainment for sports fans, and are an asset to their organizations. There should be measures taken to ensure the success of these young athletes and to protect their careers and elbows against injury due to overuse. With the invention of new technology, the methods to evaluate these occurrences with statistics and modeling may now be possible.

1. Data Sources

Two data sources will be necessary here: the data on the pitches and pitchers, and the injuries data.

* 1. Pitching Data: [pybaseball](https://github.com/jldbc/pybaseball)

Pybaseball is a python package that scrapes PITCH f/x and Statcast data from various baseball statistics websites. The package is more fleshed out than any in R that I found. I will use this to grab the data and then load it into R.

* 1. Injury Data: [Tommy John Surgery List](https://docs.google.com/spreadsheets/d/1gQujXQQGOVNaiuwSN680Hq-FDVsCwvN-3AazykOBON0/edit#gid=0)

This document has been maintained for several years by Jon Roegle, a baseball analyst and writer for The Hardball times.

1. Methods

Based on the data available for the explanatory variable, I see a few options for models for evaluating my empirical question.

* 1. Logistic Regression
     1. Building a model to Yes/No for a particular season, or seasons, whether or not the pitcher in question required Tommy John Surgery.
  2. General Linear Regression: Survival Analysis
     1. Using a GLR, or a variation to train towards number of innings pitched before injury, in order to evaluate a continuous variable.
     2. I could also train towards date in which they were injured. The number of days in the season that had elapsed prior to injury, or the number of games that had been played.

1. Preliminary Results
   1. The data for this problem has proven more cumbersome to obtain than I anticipated. The pitch level data is quite vast and to get years of data, the query can take quite a while. Then it required aggregating up to the year level for each player to give a year over year view into their professional career. The Injury list then had to be joined with the overall pitch data.
   2. Since the pitching injury provides a date of surgery, it will take some manipulation to find the season prior to the injury and find the last game played to get a view of the pitcher.
   3. A timeline plot of the injuries suggests that the rate of pitchers that require the Tommy John surgery is increasing.

