Presentation Write-up

Sabermetrics is the analysis of baseball games using the statistical methods. It is used to quantify in-game activities through analyzing the relevant data collected from the games. The analysis results can be used to compare performance of players or teams. It is also useful to provide answers to some interesting questions based on scientific arguments. From this DRP program of sabermetrics, I learned about multiple sources of baseball data, analyzing data using R, and logistic regression. The title of my final presentation is "Does more innings pitched by starters give a team higher chance of making the playoffs?"

The first thing I learned in this program is the sources of baseball data. There are three important baseball databases. The first one is Lahman's database, which collects awards, all-star, batting, fielding, pitching, team data. The second one is Retrosheet's database where you can find game logs and play-by-plays. The third one is Pitchf/x, which collects pitch-by-pitch data. Both Lahman and Retrosheet have data collected from 1871 to the most recent season. The difference is that Retrosheet has more detailed data which contains records of each game, while Lahman only has data for the whole season. As for Pitchf/x, the available data is collected from games starting in 2005, but contains detail of each pitch, including pitch speed, pitch type, spin rate, etc. Two most important statistical topics I learned from this program are logistic regression and more advanced use of R. Before, I only knew about the linear regression. Through this program, I learned a powerful method to predict the outcome of a binary case. Having more opportunities of programming using R, I got to know more methods and the process of analyzing data.

For the data in my final presentation, I referred to the data in Retrosheet for the starting pitchers, including the innings pitched, homeruns allowed, walks issued, and strikeouts, and to Lahman to see if a team made the playoffs. I used the data from 2012, the year the MLB introduced wild card, to 2019. I came up with a logistic model with the innings pitched by a starter to predict the chance of a team making the playoffs. The challenging part was to get the innings pitched by the starter. We must know when a starter was replaced, then calculated the innings that he pitched, and recorded the team that he pitched for. This scheme must be carried out for all games. Eventually, I derived the logistic model and obtained the formula $p(playoff) = \frac{exp(-9.44 + 0.00948 \times starterIP)}{1+exp(-9.44 + 0.00948 \times starterIP)}$. The formula demonstrates that the team has a higher possibility of making the playoffs if more innings pitched by the starters. Analyzing the data from Retrosheet, the shortest innings pitched by a team's starters is 624, which predicts a team having 2.8% chance of making the playoffs. On the other hand, the longest is 1033 innings, which predicts a team having 58.7% chance of making the playoffs.

I created another model to account for the pitchers' ability, called FIP. It is a combination of homeruns allowed, walks issued, and strikeouts divided by innings pitched. The lower the number is, the better the starter. The model shows that there is little to none difference between starters' innings pitched and the chance of making the playoffs.

Under the supervision of my mentor, Michael, I learned how to analyze the data, to derive the model, and to explain the outcome. The hardest and, perhaps, the sweetest part was to produce reasonable predictions. It took me a couple of weeks of trying and failing. When I was stuck, I looked for Michael for suggestions. Another lesson I learned was "Graphs are mightier than data". I am looking forward to doing similar projects analyzing data from professional sports.