Major League Baseball has been America’s sports pastime for over 100 years and was first founded in 1871 for the National League and 1901 for the American League. Today, there are 15 teams in each league. In the year 2000, the two leagues merged into what is now known as Major League Baseball (MLB). Since the very beginning, statistics in baseball have played a major role in the game. In today’s baseball, advanced metrics are being used by every major league team in order to gain advantage over their competition. MLB organizations employ data science teams to collect this information for executives, general managers and coaches. But, statistics in baseball have always been polarizing. Some managers have lost their jobs recently because they could not adopt and did not believe in advanced metrics.

Quote from Bobby Bragan (baseball manager – 1940’s) – “Say you were standing with one foot in the oven and one foot in the ice bucket. According to the percentage people, you should be perfectly comfortable.”

Quote from Leonard Koppett (A Thinking Mans Guide to Baseball – 1967) – “Statistics are the lifeblood of baseball. In no other sport are so many available and studied so assiduously by participants and fans. Much of the game’s appeal, as a conversation piece, lies in the opportunity the fans get to backup up opinions and arguments with convincing figures, and it is entirely possible that more American boys have mastered long division by dealing with batting averages than in any other way.”

Quote from Harold Reynolds (ESPN Analyst – 2004) – “I don’t understand. All of the sudden, it’s not just BA and Runs Scored, it’s OBA. And what is with O-P-S?”

Baseball contracts involve intense negotiations and millions of dollars are at stake. Recently, the St. Louis Cardinals signed Paul Goldschmidt, agreeing to a five-year, $130 million contract. Paul Goldschmidt is 31 years old. Age plays a big role in a baseball players performance. At some point as players get older, their performance on the field inevitably starts to decline. I have wondered whether the Paul Goldschmidt deal was good for the St. Louis Cardinals, and some have said, according to Forbes Magazine, there is reason to believe he could be entering the decline phase of his career. The Cardinals are betting that he will produce through the age of 37 years old. What are the chances that Paul Goldschmidt will continue to perform until the age of 37? I would like to do analysis on MLB hitters and look at past performance and predict when a batter’s performance will start to decline. My customer in this analysis is baseball teams, analysts and fans.

The “steroid era” in baseball was a dark time for MLB baseball. Even today some of the greatest players of all time are not in MLB’s Hall of Fame because of suspected steroid use. Barry Bonds is one such example. The steroid era in baseball is defined to be from 1993 to 2002. I would like to do analysis of hitting performance of the pre-steroid era (1973 to 1992), the steroid era and post-steroid era (2003 to current). I would like to know if modern advanced metrics show a significant advantage to players performance and endurance (ability to perform longer). My customer in this part of the project is analysts and fans who refuse to let players who took (or suspected of taking) performance enhancing drugs (PEDs) into the hall of fame because of unfair advantage. Did they really have an advantage?

As Harold Reynolds said, “All of the sudden, it’s not just BA and Runs Scored, it’s OBA. And what is O-P-S?” Certainly, the “old standard” hitting metrics like batting average and runs scored have given way to more advanced metrics such as OPS (on-base plus slugging) which is a more meaningful metric on how well a player is performing at the plate. OPS is calculated by adding a player’s on-base percentage with their slugging percentage.

The details of the equations are as follows:

OPS = OBP + SLG

OBP = (H + BB + HBP) / (AB + BB + SF + HBP)

SLG = TB / AB

Where:

H – total number of hits of a player

BB – total number of walks (base on balls) of a player

HBP – total number of times the player was hit by a pitch

AB – total number of plate appearances (times at bat) by the player

SF – total number of sacrifice flies of a player

TB – is the total bases and is a weighted sum ( 1 for single, 2 for double, 3 for triple, 4 for HR).

So, TB is ∑ ( (nSingle \* 1) + (nDoubles \* 2) + (nTriple \* 3) + (nHomeRuns \* 4) ) where nSingles is the number of singles, nDoubles is the number of doubles, nTriple is the number of triples and nHomeRuns is the number of home runs.

OPS – on-base plus slugging

OBP – on-base percentage

SLG – slugging percentage

NOTE: all statistics are taken over a period of time (typically a year)

(source : Wikipedia)

I would like to use OPS for both parts (performance/age prediction and PED usage) of analysis. In order to do this analysis, I have done some internet research and found raw data collected from 1871 to 2018 of major league baseball games. All of the above atomic data elements such as hits, at bats, etc are available and therefore OPS, OBP and SLG can be computed. Thanks to Sean Lahman and others, they have created a database with yearly baseball statistics from 1871 to 2018. The database has copyright 1996-2018 by Sean Lahman. I have read the license agreement which is licensed under Creative Commons Attribution and will not restrict me from using this data. The raw data needed will be from the 1960 to 2018.

Tony La Russa (ex St. Louis Cardinal Manager) has been quoted as saying (paraphrased) “you may not agree with me, but you don’t have all of the information that I have”. Now we do.

Deliverables will be code, Jupiter Notebook and PPT slide deck showing results of my work.