Moneyball Week 3 - Assignment Overview

In week 3 we used several seasons of data to complete a more robust analysis of the impact of Moneyball on the valuation of performance metrics in baseball.  Our findings from week 2 appear to be confirmed as teams appeared to undervalue OBP in the labor market before the publication of Moneyball and increased their valuation of OBP after Moneyball was published.  We find that the Moneyball has had an effect on the valuation of key performance metrics in the long run.

The assignment for this week is to replace the metrics used in our salary regression with the related metrics used in the follow up study by Hakes and Sauer (2007).  These metrics include the following:

* Batting Average = Hits/At-Bats
* Isolated Power = Slugging Percentage – Batting Average
* Eye = (Walks + Hit By Pitches)/Plate Appearances

**Beware:** Even though your code might get you to the correct answer at a given point, it is sometimes possible that the way you write it might interfere with completing a further step. So even if you get the answer right, you should look at the code we supply to check if you are going the same way. In practice, there are often many ways to get to answer in Python, and we do not insist that you follow our approach exactly – but simply warn you to be aware that differences could turn out to be problematic later.

# Assignment - Part 1

## **I. Data Preparation**

1. Import packages pandas, matplotlib.pyplot, numpy and statsmodels.formula.api
2. Read in “Master.csv” data (created from Week 2)
3. Create “Experience Squared” variable
4. Calculate variables for batting average, isolated power, and eye as defined above (for “eye” do not include IBB in walks)
5. Subset the data to only include seasons (SalYear) 1995-2015

## **II. Running Regressions for Each Season**

1. Write a function to run the following regression for every season in the data: lnSal on batting average, isolated power, eye, plate appearances, experience, experience-squared, and positional fixed effects for free agents only
2. Create a list to store regression results (note: index will have to be slightly adjusted to account for shorter timeframe)
3. Run regression for each season in the data and store results
4. Give each regression result a name which is the season it corresponds to
5. Divide results into following eras:

* Pre-MB: 1995-2000
* Moneyball Era: 2001-2008
* Post-MB: 2009-2014

6. Generate result tables for each era (should contain regression results for each individual season within each era).  Make sure to include the R-squared and number of observations in the regression output.

# Assignment - Part 3

## **III. Running the Pooled Regression**

1. Subset data to only include free agent players and add a “PreMB” dummy variable to your dataframe for seasons prior to 2004.
2. Run a pooled regression for free agent players using the format same model structure developed in your function in 6).  Your pooled regression should have the form: lnSal ~ (model from 6)) + Pre-MB\*(model from 6))