Moneyball Week 2 - Assignment Overview

In week 2, we used five seasons of data from Sean Lahman’s Database to reproduce Table 3 from the Moneyball paper of Hakes and Sauer. We show that teams appeared to undervalue OBP in the labor market before the publication of Moneyball and increased their valuation of OBP after Moneyball was published.

The assignment for this week is to include batting average in our player salary regression model. Remember that batting average is defined simply as the number of hits divided by the number of at bats. Both hits and at bats are included in the Retro sheet data.

To complete the assignment you will need to repeat the same kinds of steps we took to generate Table 3.This assignment will include three checkpoints where you will be asked to answer multiple choice questions based on your results. After answering each question you will be shown code that can produce the correct answer, which you should use as the basis for moving onto the next question.

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. Building the Dataset – Player Stats**

1. Load the “Salaries” file, drop any missing values, create a variable for the natural log (ln) of player salary, and rename the column “yearID” as “SalYear”.
2. Create a copy of the “Salaries” dataframe called “Master”.
3. Load the batting data and sum data across stints.
4. Subset batting data to only include batting seasons (yearID) 1998-2006 and players with at least 130AB.
5. Calculate PA, OBP, SLG, and batting average
6. Create SalYear variable to create one year lag between batting performance and salary
7. Merge batting data and master data.

# Assignment - Part 2

## **II. Building the Dataset – Player Info**

1. Read in “People” data and extract the player’s debut year
2. Merge debut year into master data and calculate years of experience
3. Based on a player’s years of experience, create indicator variables for arbitration eligible players (3-6 years) and free agent players (more than 6 years)
4. Read in the data for player appearances and group by stint.  Then identify the maximum number of games played at a given position for each year.
5. Create a function to determine player position.
6. Exclude non-position players.
7. Create an indicator variable for catcher and the infield (2B, SS, 3B) positions **separately.** Thus, you should have a separate indicator variable for 2B, SS, and 3B individually as opposed to one infielder indicator variable combining these positions.
8. Merge this into your master data.

## **III. Running Regressions**

Run the following regression models:

1. lnSal on OBP, SLG, batting average, plate appearances, arbitration (dummy), free agent (dummy), and all positional dummy variables during the seasons prior to the publication of Moneyball (1999-2003) combined.
2. Repeat step 2) but run the regression for the seasons 2004-2006 (all years combined).
3. Run the same regression model as in steps 2) and 3) separately for each season.  It may be easiest to read output if you display your results in a couple of tables (one for Pre-Moneyball and one for Post-Moneyball).