Louis Kabelka
3/24/2022
Statistical Computing

Baseball Report

Sabermetrics is the statistics of baseball. This project seeks to explore some of the possible analysis in this field. We will use SQL queries. The database is Sean Lahman's Baseball SQLite database.

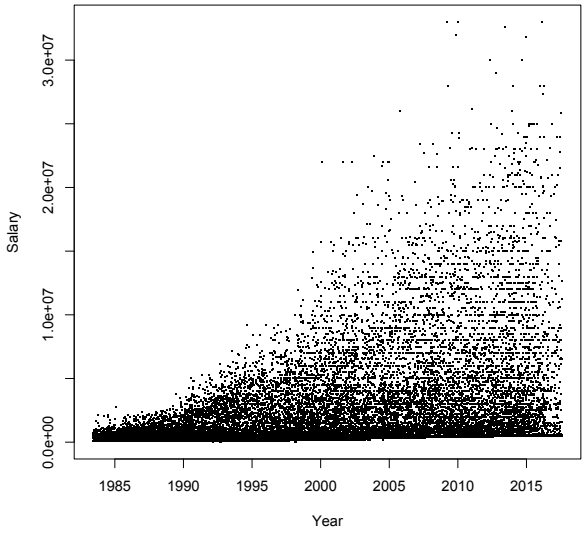
We will collect some cursory information about the database. Then we will dive deeper into discovering the causal relationship between different tables of data. There will be an emphasis on discovering the factors that effect the salaries of the various teams and players.

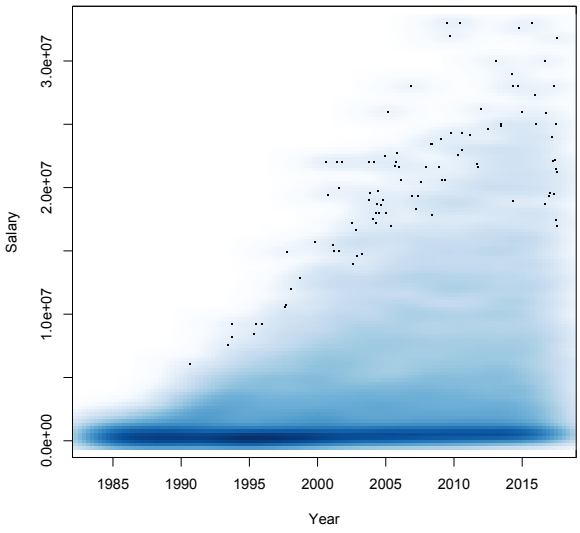
1. Number of observations with salary information. Range of years.

There are 26428 observations with salary information. Note that this does include some information of the **same players in different years**.
The range of years is 1985 to 2016.

2. Scatterplot: Salaries vs Year. // 3. smoothScater() version

The following plots show the salaries of every baseball player from every year vs the year that they played. The smoothScatter() version emphasizes the quadratic gradient of the salary distribution.





4. Salary linear regression

A linear regression to predict the salary of a baseball player based on the year they played in and their league.

(Intercept): -271424769
lIDNL: -167213
yearID: 136728

The **yearID** coefficient shows that on average between 1985 to 2016 the average salary rose by 136728$ per year.
The **lgIDNL** coefficient shows that on the average salary in the National League is 1672213$ less than in the American League.
The **Intercept** is illogical in this case, since it's estimating what the salary would be at year 0 in the National League, which we have no data for.

5. Log of salary linear regression

The logarithm of the previous linear regression.

(Intercept): -130.26534
lgIDNL: -0.04955
yearID: 0.07190

The logarithm changes the interpretation of the change in average salary from units to percentage. This means we no longer see by how many dollars the average salary changes, but by how much percent the average salary changes.

The **yearID** coefficient shows that on average between 1985 to 2016 the average salary rose by 7.19% per year.
The **lgIDNL** coefficient shows that the average salary in the National League is 0.04955% less than in the American League.
Again, the **Intercept** is illogical, since we have no data at year 0.

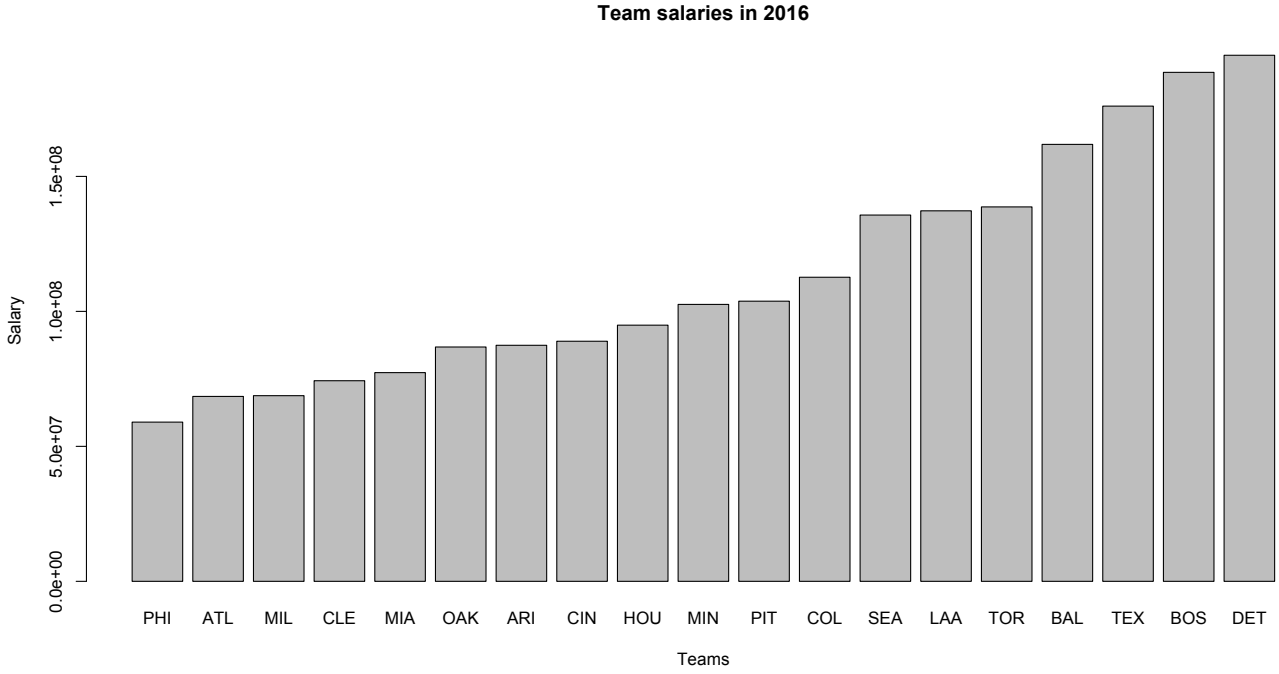
6. No log fit vs. log fit

Which of the previous two regressions fit the data better?

The R^2 values of the non-logarithmic fit is 0.1243, whereas the R^2 value of the logarithmic fit is 0.2099. A higher R^2 value indicates a better fit. Therefore, the logarithmic fit is a better fit.

7. Team salaries in 2016

The following is a bar plot of all the team total salaries in 2016.
The team with the highest salary was PHI (Philadelphia Phillies) at 58,980,000.
The team with the lowest salary was DET (Detroit Tigers) at 194,876,481.



8. Total salary of each team in each year

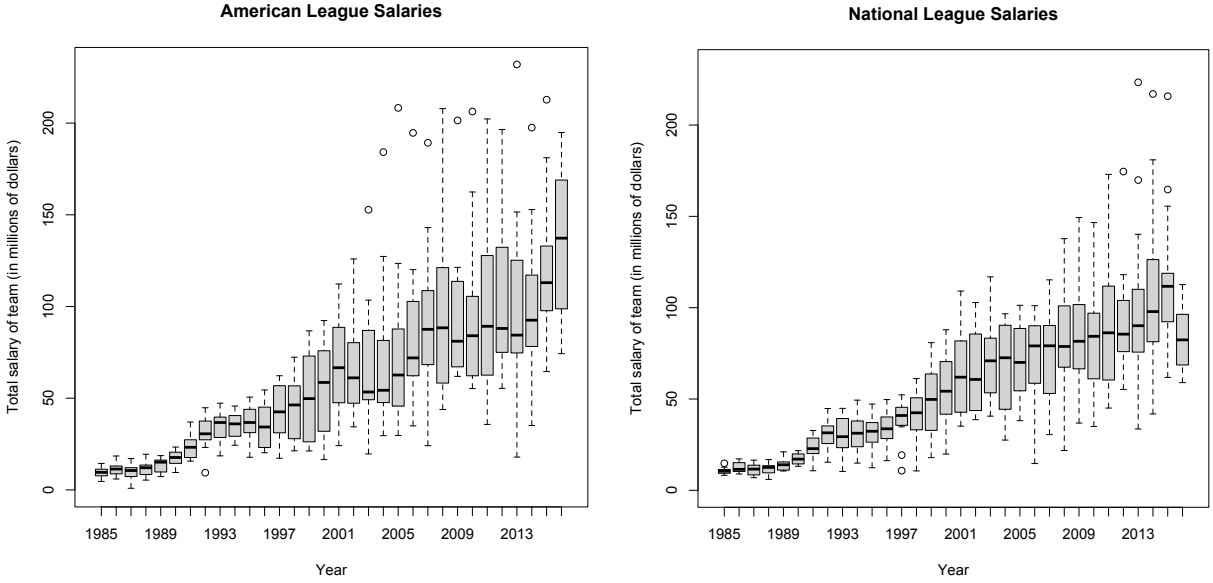
The previous query showed us that in 2016 there were 19 teams with salary information.
However, how many teams is there total salary information for in every year?

There are 907 teams with salary information.

9. Boxplot of each league

The American League and National League team total salaries over the years. Each year has a boxplot of the teams in each league.

Both show general trends in rising salaries, but also more disparity in each league. The American league seems to have slightly higher disparity and also more wider boxes, so more disparity.

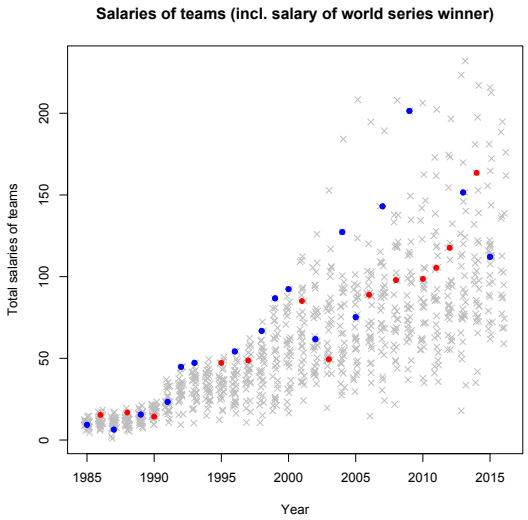


10. World Series winner

There were 17 winners in the American League. There were 13 winners in the National League.
The mean total salary of the winner in the American League is 77,596,664.
The mean total salary of the winner in the National League is 73,004,014.

This means are definitely influenced by the changing total salary of the teams over the years. In 1985, those salaries would have been way too high for any team. In 2016, those salaries would maybe be seen on the lowest paid team. So, it is more an aggregate of a well-paid team through the years.

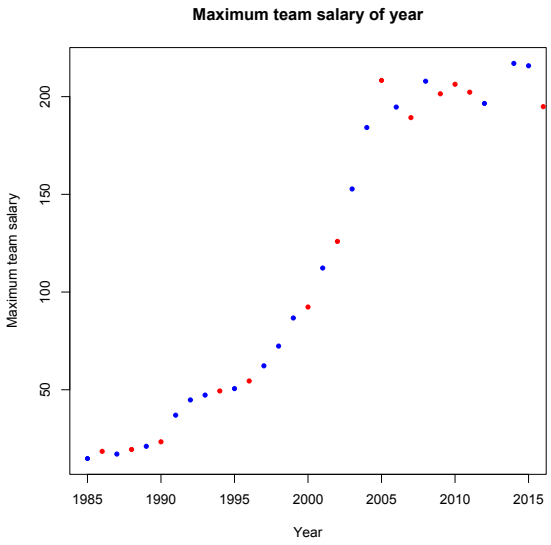
11. World Series winner vs other teams



The total team salary of the World Series winner of every year is shown.
If the winner is in the American League, it's blue. If the winner is in the National League, it's red. The salaries of all the other teams are shown as grey crosses.

In general, the World Series winner seems to be on the more well-paid side. The American League in particular often has one of its highest paid teams as the world series winner. On the other hand, the National League has some teams that have more middling pay win the world series. The early years between 1985-1991 definitely had a lot of teams that were less well-paid winning the World Series.

12. Maximum team salary of every year



The maximum total team salary for every year is shown.

In general, the maximum team salary has increased over time, but in recent years, 2005-2016, has started to stagnate. There exists a pattern, where the American League holds the maximum salary for 1-3 consecutive years and is then replaced by the National League for 1 year. This pattern has started to disappear from 2005-2016, where the National League has held the maximum salary for 3 consecutive years. It seems that the American League is no longer interested in competing for the maximum salary by upping its pay.

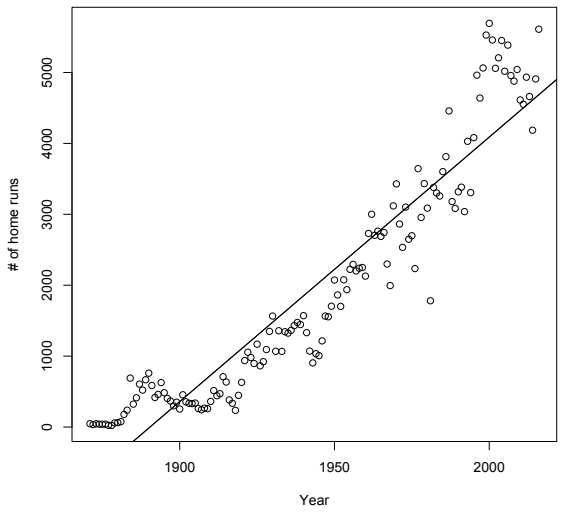
13. All-Stars winning the World Series

These are the teams that won the World Series with the most All-Stars.

In 1958, NYA won with 9 All-Stars.
In 1939, NYA won with 10 All-Stars.
In 1962, NYA won with 13 All-Stars.
In 1961, NYA won with 14 All-Stars.
In 1960, PIT won with 16 All-Stars.

14. Homeruns

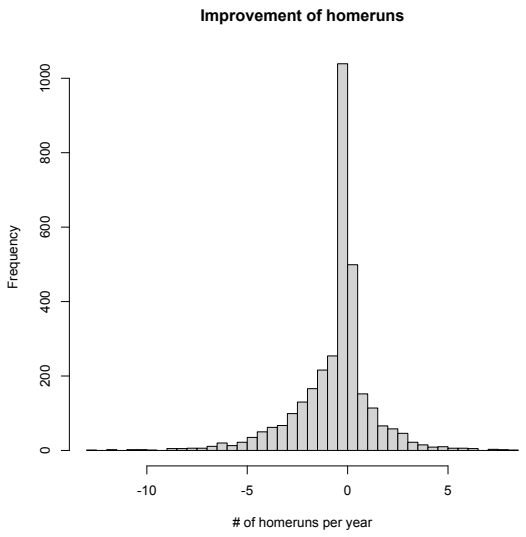
Has the total number of home runs hits by players increased over time?



We perform a linear regression on the total number of home runs from every year.

The total number of homeruns has been increasing at an average rate of 37.301 more homeruns per year.

How have individual players improved their homeruns? Only considering players with 10 or more years of batting experience.



For each player, we performed a linear regression on the # of home runs from every year.
Then we extract the slope of every regression into a dataset. The histogram shows all the slope values (# of homeruns per year) from the regression.

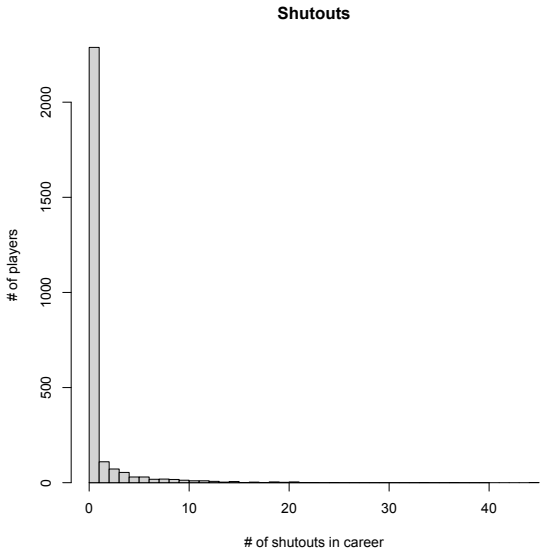
The mean is -0.5781 with Q1 being -1.2594 and Q3 being 0.0740. So, the majority of players actually score less homeruns over the course of their career.

15. Shutouts and maximum pay

For the average pitcher, is there a correlation between the total number of shutouts a player has achieved and how much they're paid at the peak of their career? What about for the non-average player?

First, we must discover what the "average pitcher" is. We will do this by judging how many shutouts most pitchers make over the course of their career.
For each player, we'll extract the # of shutouts.

```
## for each player, # of shutouts & max pay
shutouts <- dbGetQuery(con, "
SELECT SUM( Pitching.SHO ) AS shutouts, Pitching.playerID, MAX( Salaries.salary ) AS average_salary
FROM Pitching JOIN Salaries
WHERE Pitching.playerID = Salaries.playerID AND Pitching.yearID = Salaries.yearID
GROUP BY Pitching.playerID
ORDER BY shutouts
")
```

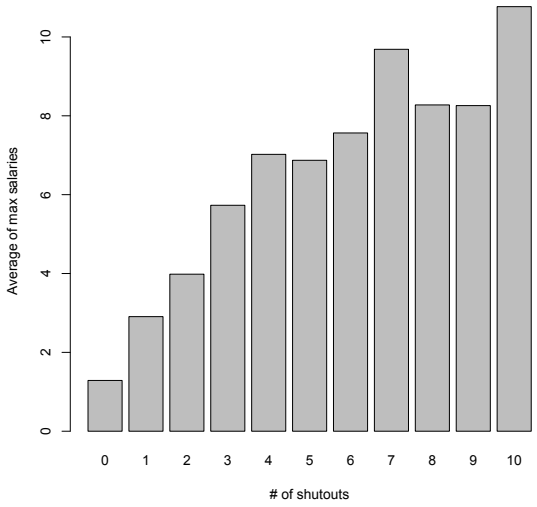


1. The histogram shows the frequency of the total number of shutouts throughout a pitcher's career. We can see a heavy right skew. It is clear that most pitchers achieve 0 shutouts. Very few ever achieve more than 10 shutouts

98.25440% of pitchers have <=10 shutouts. We will consider this the cut-off point for the average pitcher.

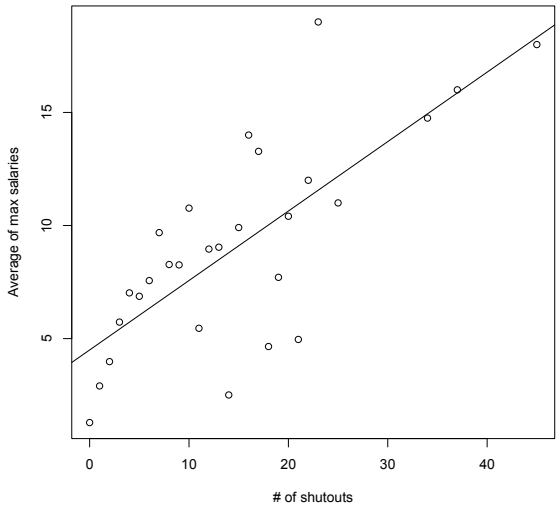
Next, we need to find out how the # of shutouts is related to the maximum a player is paid throughout their career. We'll reuse the query from before as a sub-query. Then we'll average the maximum salary for each number of shutouts.

```
## for each # of shutouts, average max pay
totalShutouts <- dbGetQuery(con, "
SELECT shutouts, AVG( average_salary ) AS average_salary
FROM (
SELECT SUM( Pitching.SHO ) AS shutouts, Pitching.playerID, MAX( Salaries.salary ) AS average_salary
FROM Pitching JOIN Salaries
WHERE Pitching.playerID = Salaries.playerID AND Pitching.yearID = Salaries.yearID
GROUP BY Pitching.playerID
ORDER BY shutouts
)
GROUP BY shutouts
")
```



2. The barplot shows the max salary for pitchers with 0 to 10 shutouts. It is clear that most pitchers that have been paid a >6 million have had at least 4 shutouts throughout their career. On the other hand, pitchers with 0 shutouts are generally hardly paid over 1 million.

How about for players that have >10 shutouts throughout their career?



3. The plot shows the increase in average max salary for each shutout a player does in their career. Pitchers that achieve over 10 shutouts are definitely over-performers. Overall, they are also well-compensated. There is a $307,100 increase in average max salary for each shutout.