An Analysis of Baseball Statistics

How Baseball Statistics are Correlated with Attendance and Team Salary

Jay Dahlstrom & Andy Siegel

3/15/2013

Overview of Research Questions

1. What is the association between various baseball statistics and fan attendance? For example, do high home run counts correlate with high attendance? We will compute the correlation between the statistics and attendance to determine which statistic has the highest correlation. In particular, we will perform two comparisons based on 2012 data: one for batting statistics and the other for pitching statistics. Furthermore, we will graph the highest correlations for batting and pitching with scatter plots.

   Based on the analysis, we determined that batting average is the highest correlated batting statistic with attendance and runs per game is the highest correlated pitching statistic with attendance.

2. What is the association between team salary and baseball statistics? Do teams that have higher salaries produce better statistics? We will compute the correlation between team salaries and various statistics to determine which statistics have association with salaries. Analysis will focus on batting, pitching and salary data from 2012. In addition, we will graph the highest correlations for batting and pitching with scatter plots.

   Based on the analysis, we determined that batting average is the highest correlated batting statistic with team salary and strikeouts is the highest correlated pitching statistic with team salary.

Motivations and background

Professional sports are a big business in the United States. Each year, millions of fans attend games and teams spend millions of dollars to hire the best players. Baseball franchises make a significant amount of their revenue on ticket sales and concessions. For franchises to remain economically viable, it is important for them to attract as many people as they can. Therefore, the financial stability of the franchise relies on fan attendance levels. The first part of our analysis explores the correlations between various baseball statistics to attendance. These correlations might be useful for showing the impact each statistic has on attendance levels. For example, are home runs more associated with attendance than runs batted in? The answers to these questions could support studies conducted by team marketing strategists. Another important association to consider is between team salary and statistics. Teams spend large sums of money to bring in players with the best talent. Do teams that spend more money on players produce better statistics than teams with lower salaries? Results gathered from this analysis might be important for baseball managers, owners and scouts. Spending large amounts of money on players is a significant investment and it is important for franchises that this investment pays off in the long term. Answers derived this analysis might suggest how salary is associated with various statistics. Overall, this analysis looks at human preferences, ranging from what makes fans attend games to what makes teams spend large amounts of money on player salaries.

Dataset

The primary dataset used in this analysis is baseball-reference.com. This website contains every imaginable statistic in recorded baseball history. For the purposes of this project, the focus is on the data from the 2012 season. The dataset contains information on batting and pitching (http://www.baseball-reference.com/leagues/MLB/2012.shtm), and miscellaneous information (http://www.baseball-reference.com/leagues/MLB/2012-misc.shtml). In both cases the data is organized by team. The batting and pitching dataset contains information such as the number of Home Runs a team hit in a year, the team's earned run average, and the number of strikeouts. On the other hand, the miscellaneous dataset contains information on team salary, attendance, name of the general manager and other similar data. This information will form the basis of the analysis.

The second dataset is wikipedia, which was used only to acquire the capacity of all the stadiums in the MLB. This information is available at the following web address http://en.wikipedia.org/wiki/List_of_Major_League_Baseball_stadiums. To answer the research question it is necessary to know the capacity of every stadium so that the analysis can be uniform for each team. Since different stadiums have different capacities, the results would be biased if we compute attendance purely on the number of people who attended games. Therefore, percentages are the only way to make comparisons.

Instructions for downloading data:

- for batting and pitching data:
    1. go to http://www.baseball-reference.com/leagues/MLB/2012.shtml
    2. copy all of the rows until the row named WSN under 'Team and League Standard Batting', thus excluding the last three rows
    3. paste the rows into an excel spreadsheet
    4. repeat step 2 for the data in 'Team and League Standard Pitching" and paste the results at the end of the same excel file mentioned in step 3
    5. save the file as baseball-statistics.csv in the working directory
- for the miscellaneous data
    1. go to http://www.baseball-reference.com/leagues/MLB/2012-misc.shtml
    2. copy all of the rows under 'Miscellaneous Team Info'
    3. paste the rows at the end of baseball-statistics.csv
    4. save the file
- for capacity data:
    1. go to http://en.wikipedia.org/wiki/List_of_Major_League_Baseball_stadiums
    2. sort the table under 'Current Stadiums' by the Team row (this will sort the data in alphabetical order the same manner as baseball-reference.com)
    3. copy only the data from Seating Capacity row and appended it to baseball-statistics.csv
    4. Change column name from Seating Capacity to Capacity
    5. save the file

Methodology (algorithm and analysis)

The goal of our analysis was to identify the batting and pitching statistics that have the highest correlation with attendance and team salaries respectively. To complete this analysis, we wrote a python program to compute the correlations. The particular statistics analyzed included batting and pitching statistics. For batting, we selected runs, hits, home runs, stolen bases, and batting average. For pitching, we selected runs per game, win/loss percentage, the number of complete games, earned run average, and strikeouts.

For each of the selected statistics in the two groups (batting and pitching), we calculated the Pearson's r (correlation coefficient) with attendance and salary, respectively. The results produced by Pearson's correlation test were numeric values between -1.0 and 1.0, where -1.0 was perfect negative correlation, 1.0 was perfect positive correlation and 0.0 was no correlation. The correlation coefficient allowed for comparisons across the different statistics to see which had the greatest association with attendance and salary. Furthermore, we graphed four scatterplots corresponding with the highest correlations.

Results

Batting average was the highest correlated batting statistic with attendance. The correlation value was .5257. This means that there was a relatively strong positive correlation between batting average and attendance in 2012. It is important to note that the correlation strength does not specify causation. The correlation value simply determines the degree of association between the two variables. Although this correlation is not incredibly strong, it does exhibit a relatively strong association. Therefore, MLB marketing strategists might further investigate relationship between batting average and attendance to determine how they might influence each other. On the other hand, runs per game was the highest correlated pitching statistic with attendance. The correlation value was .4514. This value indicates a moderate positive correlation for the 2012 season. While this value is not as strong as the value for batting average, there is still somewhat of an association. It might be useful for MLB marketing strategists to research the relationship between runs per game and attendance to explore how they affect one another.

Batting average was also the highest correlated batting statistic with team salary. The correlation value was .4849. This means that there was moderate positive correlation during the 2012 season. The scatter plot provided visual evidence of this finding, as the points were moderately clustered in a positive linear pattern. This result suggests that it might be important for team scouts to study the relationship between batting average and salary. In this report, we determined correlation, but if MLB teams determine causation between the two variables, then the results would be integral for player acquisition strategies. The number of strikeouts was the highest correlated pitching statistic with team salary. The correlation value was -.5332. This indicates that there was a relatively strong negative correlation during the 2012 season. The result from this computation is surprising because as the team salary increased, the number of strikeouts decreased. One would expect this to be a positive correlation, but that is not the case for this correlation test. MLB owners might conduct further research to determine if there is causation between the two variables.

Reproducing the results

To reproduce the results of our analysis, follow these directions:

1. Download the data as explained in the Datasets section of the report
2. Select the following variables and statistics for the analysis:
   Variables → 'Attend/G', 'Capacity', 'Est. Payroll'
   Statistics → 'R', 'H', 'HR', 'SB', 'BA', 'R/G', 'W-L%', 'ERA', 'CG', 'SO'
3. Extract, read and clean the data (the end result of this process is the data should be floats)
4. Compute the attendance percentage by dividing the corresponding values from Attend/G by Capacity
5. Pair each statistic with its corresponding attendance percentage and payroll (aka salary). This should result in pairs of stat-attendance and stat-payroll.
6. Sort the pairs by statistic from best to worst. All of the statistics except ERA should be sorted in descending order because higher values are better for those statistics. ERA should be sorted in ascending order because a lower vale for ERA is better.
7. Calculate the Pearson's correlation test for each statistic-variable pair. For example, the test will compare 30 attendance values with 30 homerun values to determine the correlation coefficient (r). The computation produces tow values. The value on the left is the correlation value. The results will be a numeric value between -1.0 and 1.0, where -1.0 is perfect negative correlation, 1.0 is perfect positive correlation and 0.0 is no correlation.
8. Return the highest correlations between batting-attendance, pitching-attendance, batting-salary, and pitching-salary.
9. Graph four scatter plots, one for each of the highest correlations.
10. Interpret the results using the following criteria:
    a. If Pearson's correlation value is between -.09 and .09, there is no association
    b. If Pearson's correlation value ranges from -.3 to -.1 or .1 to .3, there is a small association
    c. If Pearson's correlation value ranges from -.5 to -.3 or .3 to .5, there is a moderate association
    d. If Pearson's correlation value ranges from -1.0 to -.5 or .5 to 1.0, there is a strong association

Collaboration

No one helped us for part 2 of this assignment

Reflection

First, we learned how work together on the same program. It was a challenge deciding how to divide the workload of the code. However, through clear and constant communication, we were able to exchange updates to the code efficiently. In addition, we learned more about function decomposition. In certain sections of our code, we broke functions into several helper functions. This made the code easier to read and reuse. If we

had an opportunity to view programming projects from past quarters, we could have had a better understanding of the scope and expectations. Overall, we were satisfied how the project turned out. For future projects, we might explore various file-sharing services, such as cloud storage. During this project, we emailed each other updated versions of our code and this was occasionally tedious. We recommend selecting an interesting topic and starting as early as possible.