Measuring Correlation between Statistics and Wins in Major League Baseball

CSE 140 Homework 9

Mark Frerker

3/15/2013

1. Summary of Research Questions and Results

1. What is the correlation between strikeouts per walk (K/BB) from a pitching perspective and winning percentage? I will compute how many strikeouts a team records in a season as well as how many walks they give up. Dividing total strikeouts by total walks will give me that teams K/BB ratio for the season. I will compare that to the team's winning percentage and calculate the Pearson coefficient.
2. What is the correlation between team homeruns and winning percentage? I will compute the total number of homeruns hit by each team in the season as well as their winning percentage for the season. I will then compute the Pearson coefficient to see how much of an effect homeruns actually have on winning percentage.
3. What is the correlation between on base percentage (OBP) and winning percentage? I will calculate a team's OBP in a season, which is the total number of times a batter reached base divided by the total number of plate appearances. Using the OBP and the team's winning percentage for the season, I will calculate the Pearson coefficient to determine how much of a factor OBP plays in winning percentage.

2. Motivation and Background

Major League Baseball reported revenues of over $7.7 billion in 2012. It is a large and still growing industry that has global impact. In the past decade there has been an explosion of new methods of statistical analyses and people are coming up with new ways to evaluate players and teams all the time. The book *Moneyball* by Michael Lewis showed how a team with significantly less payroll than almost every other team in baseball, the Oakland Athletics, can use under-utilized statistics to build a winning team. The basic concept behind the moneyball strategy is to find and exploit market inefficiencies. At the time it was written, advanced statistics did not play much of a factor in the decisions made by front offices in baseball. That drastically changed in the few years following the books release, and it is now considered abnormal for a team to not employ a sabermetrician (someone knowledgeable in advanced baseball statistics). One of the key statistics mentioned in Moneyball is OBP, which is why I will be using it as one of my research questions. Some people criticized Moneyball's findings for ignoring other aspects such as Oakland's pitching, so I will be doing an elementary analysis to determine the feasibility of OBP being as important as Lewis claimed it to be. It is important to test relationships between many different statistics in order to find these hidden relationships that could lead to an improved baseball team. Knowing these answers could start a team's path from last place to playoff contender.

3. Dataset

The dataset that I will be using is a text file with game logs from every game played in the Major Leagues in 2012. The download can be found at http://www.retrosheet.org/gamelogs/index.html. To download the file, go to the previous link and click on “2012” under the “Game Logs for Individual Seasons” section. The download will be a zip file. Make sure to unzip the file into the same folder that the program is in. Each game has 161 fields, including date, score, team statistics, players, attendance, and more. The fields that I will be using are (for home team and visiting team): team, homeruns, score, strikeouts, walks, hits, hit-by-pitch, intentional walks, at bats, and sacrifice flies. A guide to the game logs can be found at: http://www.retrosheet.org/gamelogs/glfields.txt. It lists what is in each game log and the order in which the information is presented.

Example game log:

```
"20120328","0","Wed","SEA","AL",1,"OAK","AL",1,3,1,66,"N","","","","TOK01",44227,184,"00010000002
","00010000000",39,9,1,0,1,3,1,0,0,0,0,4,2,1,1,0,4,3,1,1,0,0,33,7,1,0,0,0,39,6,3,0,0,1,0,0,2,0,0,10,2,1,0,0
,7,6,3,3,0,0,33,19,1,0,1,0,"hallt901","Tom Hallion","nelsj901","Jeff Nelson","hudsm901","Marvin
Hudson","belld901","Dan Bellino","","(none)","","(none)","wedge001","Eric Wedge","melvb001","Bob
Melvin","wilht001","Tom Wilhelmsen","caria001","Andrew Carignan","leagb001","Brandon
League","ackld001","Dustin Ackley","hernf002","Felix Hernandez","mccab001","Brandon
McCarthy","figgc001","Chone Figgins",5,"ackld001","Dustin Ackley",4,"suzui001","Ichiro
Suzuki",9,"smoaj001","Justin Smoak",3,"montj003","Jesus Montero",10,"carpm001","Mike
Carp",7,"olivm001","Miguel Olivo",2,"saunm001","Michael Saunders",8,"ryanb002","Brendan
Ryan",6,"weekj001","Jemile Weeks",4,"pennc001","Cliff Pennington",6,"crisc001","Coco
Crisp",7,"smits002","Seth Smith",10,"suzuk001","Kurt Suzuki",2,"reddj001","Josh
Reddick",9,"cespy001","Yoenis Cespedes",8,"alleb001","Brandon Allen",3,"sogae001","Eric
Sogard",5,"","Y"
```

4. Methodology

First the data needs to be read and stored. Since each of the research questions involves correlations between a team’s winning percentage and some statistic, figure out all of the teams represented in the dataset. Next, compute the winning percentage for each team during the season. Winning percentage is calculated as the number of wins divided by the sum of wins and losses, since there can be no ties.

For the first research question, the number of strikeouts and walks per team over the course of the season will need to be computed. Divide the total strikeouts by total walks to get the K/BB ratio for that team. Use the Pearson coefficient equation to calculate the correlation between K/BB and winning percentage. The equation for this is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

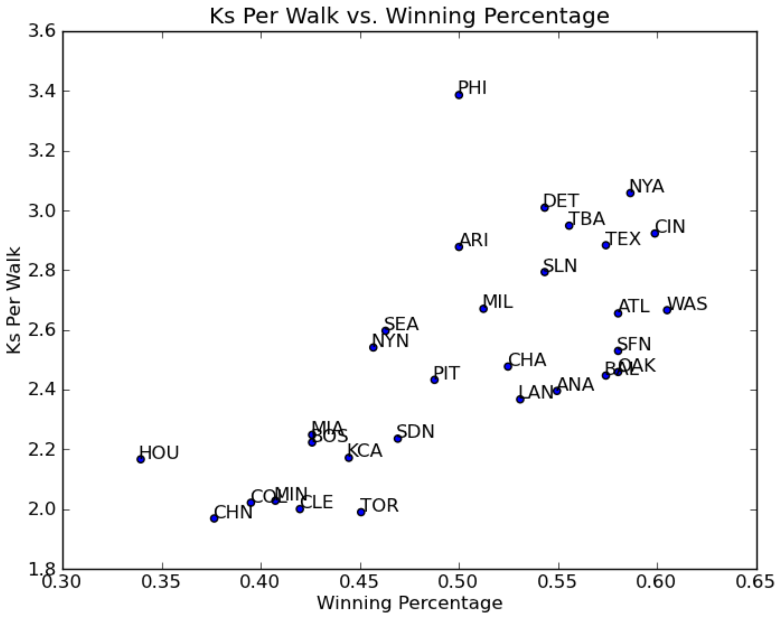
Use winning percentage for X, and K/BB ratio for Y. The resulting r should be between -1 and 1. This number represents how well a best fit line would work for this data set. An r of 1 or -1 would mean that a linear equation would fit the data perfectly, either increasing or decreasing respectively. An r of 0 would mean that there is no correlation at all between the two data sets, and changing either x or y would have no effect on the other. Anything greater than .5 or less than -.5 would be considered a high correlation. A coefficient between .3 and .5 or -.3 and -.5 would be a medium correlation. Between .1 and .3 or -.1 and -.3 is a small correlation, and anything between -.1 and .1 has effectively no correlation. For further analysis, plot the K/BB and winning percentage on a scatter plot, with winning percentages as the x-axis. Label each marker with the team's symbol (ex. "SEA" for Seattle). This will help to identify any outliers and determine their effect on the analysis. For example there could be a team that has a relatively low K/BB ratio but still manages to win a lot of games

To answer the second question, use a similar method as before, but instead using homeruns instead of K/BB ratio. The homeruns will simply be the total number of homeruns hit by a team across all of their games. Use the same winning percentage for each team as before to calculate the correlation between total homeruns and winning percentage. On a scatter plot, graph the number of homeruns hit by a team against the winning percentage for that team, again with winning percentage on the x-axis. Again look for outliers and how they might affect the data, such as a team that wins a lot of games without hitting many homeruns.

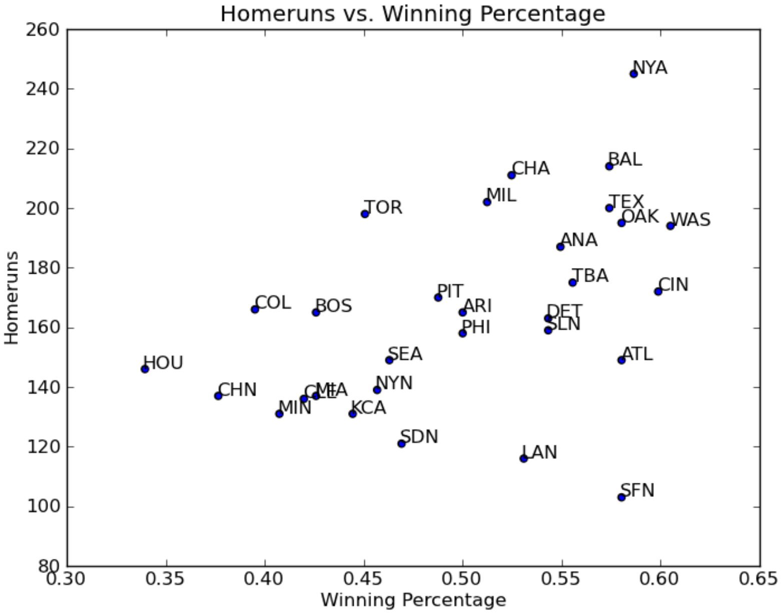
For the third research question, determine the total OBP for each team for the season. Specifically, OBP is the sum of hits, walks, and hit-by-pitches divided by the sum of at bats, walks, sacrifice flies, and hit-by-pitches. Using the same winning percentage for each team as before, compute the correlation coefficient for OBP and winning percentage. Use the same parameters as before to determine how significant the correlation is. Plot the data on a scatter plot with winning percentage on the x-axis, and look for any outliers. An outlier could be a team that loses a lot of games despite a high OBP and might be a starting point for more analysis. Compare the coefficients for each of the three research questions and analyze the differences and what could be the causes or implication of them.

5. Results

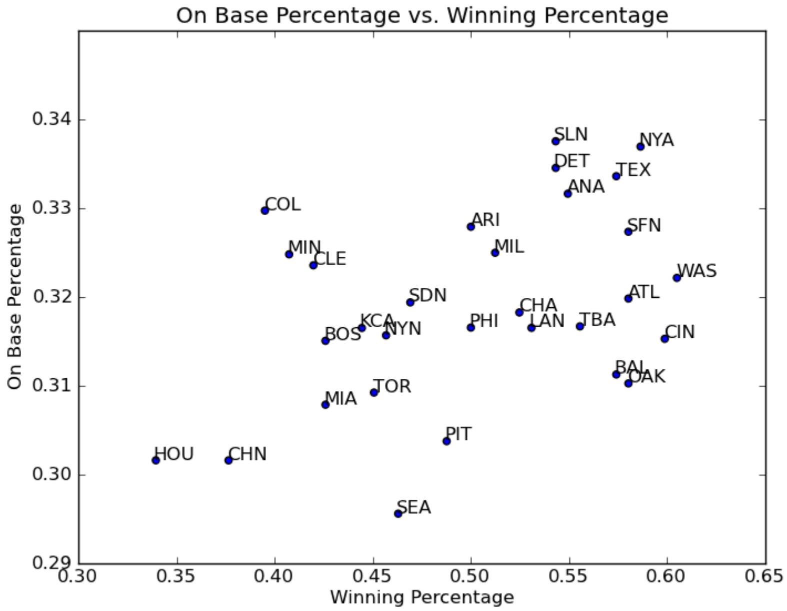
1. The correlation coefficient for K/BB ratio and winning percentage is .68116. This is a very high correlation that strongly suggests that K/BB is one of the more important factors in winning baseball games. There could be several reasons for this. One possibility is that by striking out a lot of batters, the ball is put in play less, which reduces the opportunity for errors, sacrifices, or unlucky bounces. On the other hand, it could be a result of having less opposing baserunners due to a low number of walks, which obviously makes it harder for the other team to score. The only significant outlier from the scatter plot was the Philadelphia Phillies. They had the highest K/BB ratio of all the teams but were in the middle of the pack for winning percentage. The most likely explanation for this would be a very poor offense. However it could also be the result of throwing a lot of strikes, as this would lead to a lot of strikeouts as well as a lot of hits, with a low number of walks.



2. The correlation coefficient for homeruns and winning percentage is .44884. This suggests a medium correlation, though it is close to being a high correlation. Many people would likely be surprised that there was not a higher correlation. Fans generally want their teams to get the big homerun hitters, as they believe it will give them the best opportunity to win. However there are a couple factors that I believe could lead to this lower than expected correlation. The first is obvious; offense is only half the game. While scoring runs is important, teams also need to be adept at preventing runs. Second, homeruns are generally not conducive to scoring in streaks. The average batting average with the bases empty is lower than the average batting average with at least one runner on base. Since a homerun clears the bases, it allows the pitch to then pitch from their motion, rather than pitching from the set. San Francisco was the significant outlier from this data set (high winning percentage with few homeruns), and the reason for that is fairly simple. The home stadium for San Francisco is one of the most pitcher friendly stadiums in baseball. Combine that with one of the top pitching rotations and it is easy to see how they were able to win many games without a large number of homeruns.



3. The correlation coefficient for OBP and winning percentage is .42228. This is also a medium correlation that is close to being a high correlation. It is also very similar to the correlation for homeruns, which seems to back-up the claim in Moneyball that OBP as a very important offensive statistic to consider. One of the reasons why OBP is such an important statistic is that once you get a guy on base it opens up a lot more possibilities. The defense has to play differently, which could lead to more errors or hits that would have been outs otherwise. It puts more stress on the pitcher because he has to be wary of stolen base attempts, or he may try too hard to get the next guy out and make a mistake. A reason why the correlation is not even higher is that getting on base is not enough in itself; making sure the runners actually score is still necessary. The only two outliers from the dataset can again be explained by ballpark effects. Colorado, which had a high OBP compared to winning percentage, plays in the most extreme hitters park in baseball, which would naturally elevate their hitting statistics as well as their opponents. Colorado also had a poor pitching staff, which would lead to their lower win percentage. On the opposite end is Seattle, which had a relatively low OBP compared to winning percentage. Again, this is because Seattle plays in a very pitcher friendly park that stifles the hitting statistics of both the Mariners and their opponents. Seattle also had one of the better pitching staffs in the league.



6. Reproducing Your Results

To obtain the Pearson coefficients for each of the statistics that were analyzed, enter “python baseball.py” into the command line. Make sure the current directory is the same one that the program and the data set are located in. See the Dataset section for information on downloading the data. Each of the graphs will be saved into the same folder as the program and can be viewed from there to determine outliers. Use the parameters outlined in the Method section to determine whether the correlation is large, small, or none. Additional team statistics can be found at http://espn.go.com/mlb/stats/team/_/stat/batting. For pitching statistics replace ‘batting’ from the url with ‘pitching’. These statistics can be useful in determining the cause of outliers.

7. Collaboration/Reflection

Collaboration: I did not collaborate with anyone else.

Reflection: I think this was great practice for how to do a programming project from start to finish, and has strengthened a lot of skills that will be helpful in my future jobs (even if it’s not a programming job). This was one of the more fun projects I’ve done in my academic career so far, and has even giving me ideas for other projects I could do in my spare time. It was very nice to be able to apply the things I have learned in class to something that I am very passionate about, and I’m very proud of how much I have accomplished so far this quarter. Since I was excited for this project, I think I had a very good approach to the assignment. After playing around with the dataset for a bit, I made a general plan and was able to stick to it for the most part. Even after getting feedback from part 1 I was able to implement the necessary changes easily due to my code being well laid out.