S4.3: Capstone I Proposal
By Brian Camp

Baseball Play-by-play data investigation.

Question: How do we estimate the probability of the home team winning?

This question is of interest to baseball teams in Major League Baseball and in particular, to the pitching staffs for each team. The predictions that are possibly made with this data may help teams devise pitching strategies that may aid them in winning more games during a season.

Baseball play-by-play data is available at Retrosheet.org. Using this data, a matrix of game variables will be constructed. Of particular interest is pitch sequence i.e. at a given plate appearance, what is the order of pitches thrown (B=Ball, S=Strike, X=put in play by batter) and related numbers.

As a side note, for one season of data, there are roughly 6500 rows per team x30 which is close to 200000 rows per year. The data is made available through some dos scripts which do some batch processing on the raw data that has been collected. It is possible (although not my goal in this project) that the data files could be accessed over the web and have python do the appropriate data processing (to construct the initial data frames) instead of using dos. Even with the data provided by the dos files, there are about 30 separate files that need to be processed with up to 96 possible data fields to choose from - per season. In this study, I am intending to go back to 2010 (about 9 years worth of data). So when everything is compiled together we should have a matrix with approximately 1.8million rows or so. The columns to be considered (a subset of the 96 provided and some which need to be processed to be calculated) are described below.

Matrix setup of things to consider.

Each row of the matrix will be one game
The fields in the matrix will be:

Note: home/away indicates two separate columns. Home team columns will be grouped first followed by the away team columns in the dataframe/matrix.

Batters faced (home and away) - i.e. plate appearances of other team
# of pitch sequences per game starting as: (home/away)

X
SX
SS
SB
BX
BS

BB
Total pitches thrown in game (home/away)
Total number of pitchers used in game (home/away)
Number of pitch sequences with 1-3 pitches (home/away)
Number of pitch sequences with 4-5 pitches (home/away)
Number of pitch sequences with 6+ pitches (home/away)
Number of batters allowed on base (home/away)
Number and/type of hits allowed (home/away) i.e. #1b, 2b, 3b, hr
Number of runs allowed (home/away) - i.e. # of runs the other team scores

Target variable(s):

Classification problem: Did the home team win (y/n)? or to differentiate the wins a bit futher they could be described:

Home team wins by 1
Home team wins by 2-3
Home team wins by 4+
Home team loses by 1
Home team loses by 2-3
Home team loses by 4+

Regression Problem: Track each game by the run differential (home - away = run difference). In this case this would be a regression problem.

The deliverables for this project will be the code showing the classification and regression problems and then a paper that will explain the results with graphics.