Introduction to Sabermetrics Using R

PHIL BANGAYAN

SOCAL CODE CAMP

DECEMBER 3, 2017

Overview

- Introductory talk using R applied to sabermetrics
  – No experience in either topic required

- Cover where to find data and how to process it
  – Lahman season stats to verify the Pythagorean theorem of baseball
  – Retrosheet play-by-play stats to analyze Kershaw's 2015 season
  – PITCHf/x pitch-by-pitch information to look pitch locations from Game 5 of this year's World Series

Lahman Database

- Provides statistics for each player/team by season

- Download by going to http://seanlahman.com/ baseball-archive/statistics

- Click on "2016 – comma-delimited version"

- Creates folder "baseballdatabank-2017.1"
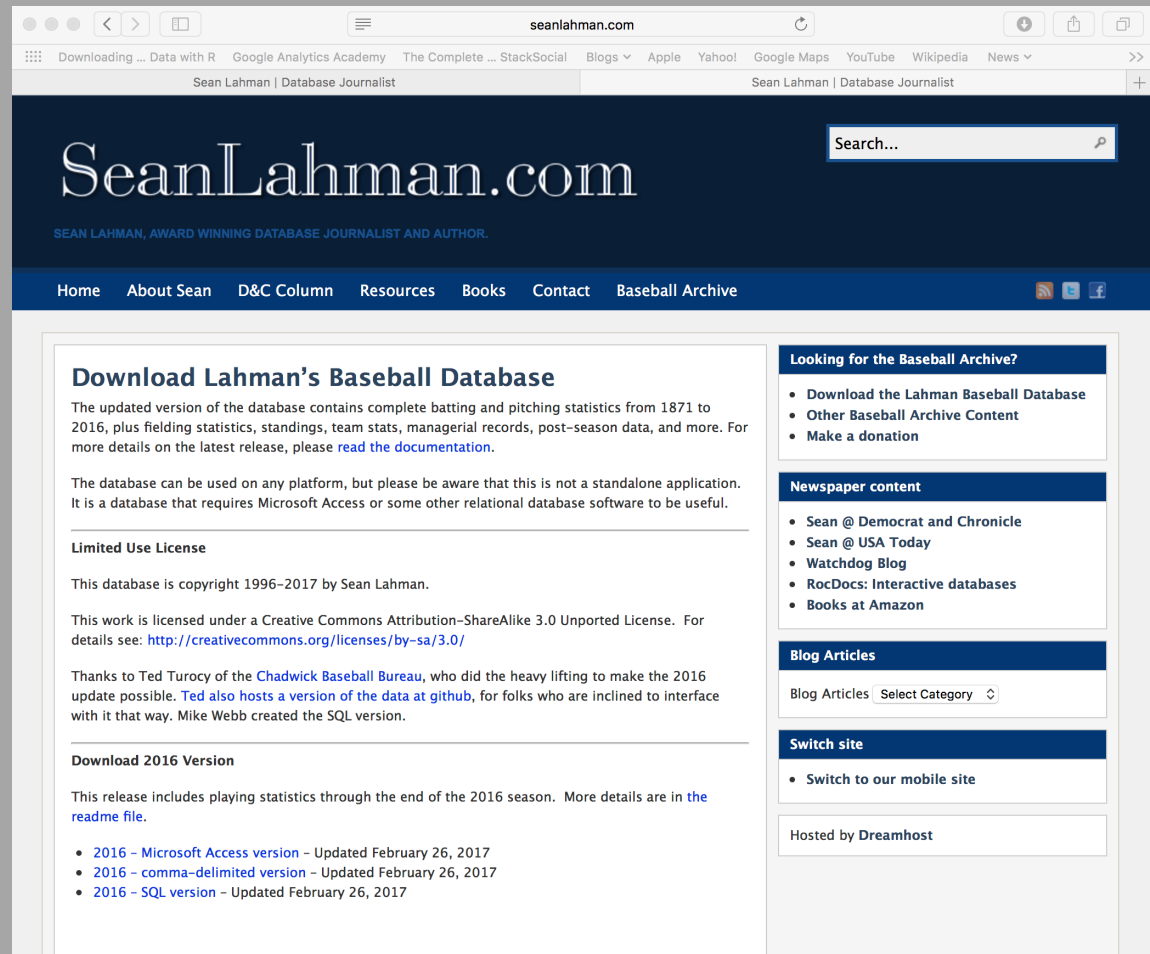
Lahman Database Tables

- Consists of 24 different tables, with main ones being:
  - Master: list of 19k different players (height, weight, birth, death, IDs)
  - Batting: offensive stats by player by team and year
  - Pitching: pitching stats by player by team and year
  - Teams: records and stats (offensive and defensive) for each team by year
- Documentation on list of tables found in readme2014.txt file
- For this exercise, will look at Teams table

Pythagorean Theorem of Baseball

- Predicts a team's winning percentage using runs scored and runs allowed

$$\text{Win Pct} = \frac{Wins}{Wins + Losses} \approx \frac{(runs\ scored)^2}{(runs\ scored)^2 + (runs\ allowed)^2}$$

- Use R and the Teams table in Lahman database to verify estimate

Analysis using R

- Copy teams.csv file into R working directory
- Set R working directory appropriately
- Read file using read.csv()
- Examine layout using str()
  - Determine useful fields as teamID, W, L, R, RA
- Limit to 2016 teams using subset()
- Calculate winning percentage, predicted winning percentage, and error
- Calculate error for entire season using mean average deviation (MAD) and root mean square error (RMSE)
- Visualize results

Resulting Plot
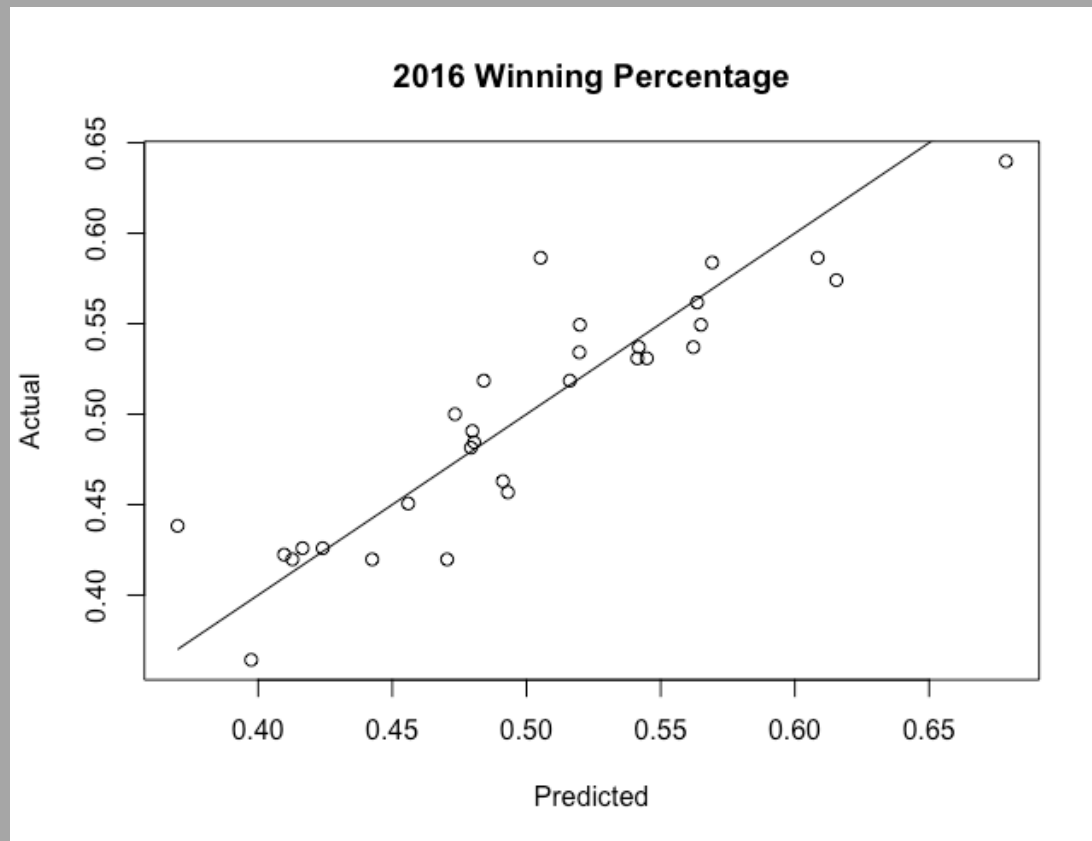
Further Analysis

- Why do some teams win more than the predicted value?
  - Determine outliers using identify() function
  - Examine effect of closers

- What is the effect of scoring an extra xx runs per season?
  - Plug back into formula, using an average value for runs scored

- Is there a better value than 2 for the exponent?
  - Solve by plotting MAD or RMSE for different exponents
  - Or solve for exponent using linear regression

Retrosheet Play-by-Play Data

- Event data for each play of each game in season
- Found at www.retrosheet.org/game.htm
- Go to "Regular Season Event Files" and download "2015" data
- Creates folder "2015eve"
    - Files sorted by home team (e.g. 2015LAN.EVN contains all Events for the National league team Los Angeles played in their home stadium)

Sample Retrosheet Event File

- id, LAN201504060
  - Identifies this section as game in Los Angeles on 2015-04-06, game 0

- info,visteam,SDN
  - Information lines for visiting team, date, time, temperature, umpires, etc.

- start,myerw001,"Wil Myers",0,1,8
  - Starting players, identification, name, visiting team, batting order, position

- play,1,0,myerw001,02,FCH,HP
  - First inning, top half, Wil Myers, on 0-2 count, foul, called strike, hit batter

Pre-processing for Retrosheet Data

- Data needs to be pre-processed before bringing into R
  - Combine all plays from all teams into one file

- Utilize other people's work
  - Chadwick software tools for scoring baseball games
    - ‣ Home page at http://chadwick.sourceforge.net/doc/index.html
    - ‣ Source at http://sourceforge.net/projects/chadwick/files/
    - ‣ Installs command line tools for parsing event files
  - Analyzing Baseball Data with R function to create CSV from event files
    - ‣ https://baseballwithr.wordpress.com/2014/02/10/downloading-retrosheet-data-and-runs-expectancy/

Pre-Processing Steps

- Install Chadwick software
    - Download source
    - Configure, Make, Make install (into /usr/local/lib and /usr/local/bin)
- Create appropriate folders and files
    - "download.folder" (in R current working directory)
    - Subfolders "unzipped" and "zipped" inside "download.folder"
    - "fields.csv" from Analyzing Baseball Data with R github site for headers
- Create data files
    - "source parse.retrosheet2.pbp.R"
    - "parse.retrosheet2.pbp(2015)"

Strikeouts in 2015 Season

- Read in data file generated in pre-processing step
- Limit play-by-play data to strikeouts by Clayton Kershaw
    - Find Kershaw's unique Player ID (kersc001, found in roster2015.csv)
    - Subset by Player ID and Event CD (3 for strikeout)
    - Extract date from unique Game ID
    - Determine cumulative strikeouts
    - Plot cumulative strikeouts by date
- Generalize into a function
- Find cumulative strikeouts for two more players
- Add to existing plot

Resulting Plot
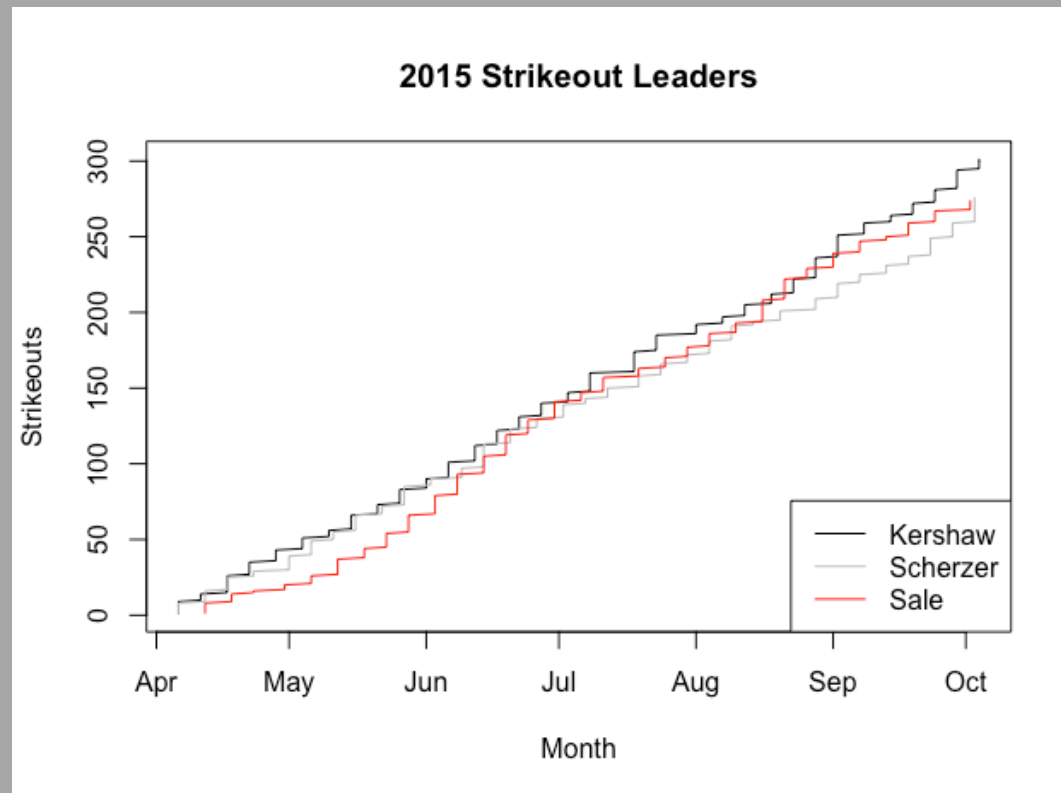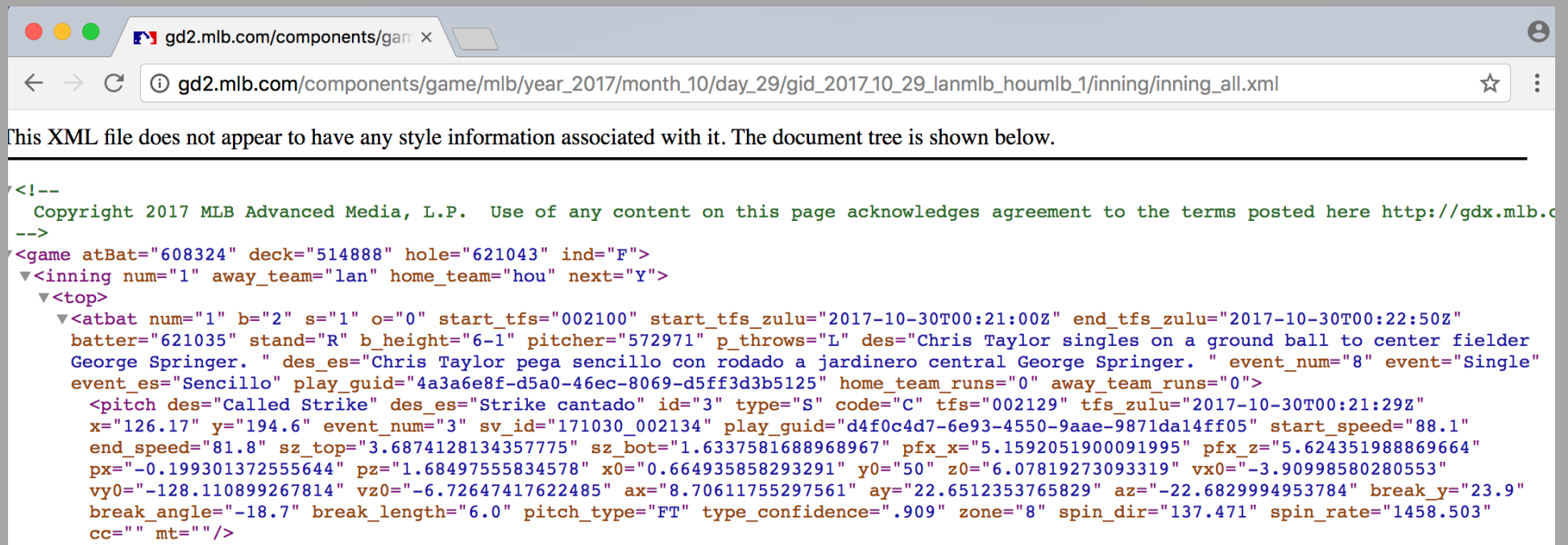
PITCHf/x Database

- Specifics of each pitch of every game since 2008
  - Type of pitch, location, velocity, acceleration, spin, hit location (if hit)
- Found at gd2.mlb.com/components/game/mlb/year_xxxx/month_xx/day_xx
  - Insert numbers for x's above
  - E.g. gd2.mlb.com/components/game/mlb/year_2017/month_10/day_29
- Leads to text webpage with links for all games that day
- Click on desired game, then click on "inning/" and then desired inning (or all)
- Results in XML page describing each pitch

Sample Data

- Just the first pitch from Game 5 of 2017 World Series

Importing Data into R

- Use XML package, particularly xmlParse()
- Helper function grabXML()
  - Published in Analyzing Baseball Data with R
  - Handles cases of incomplete data
  - Subset either to atbat, pitch, or hip attributes
- Results in data frames interpretable in R

Exercise

- Load Game 5 data (October 29, 2017)
- Focus on top of first inning pitches (n=32)
- Examine types of pitches thrown
  - table() function in R
- Plot location broken out by type of pitches
  - plot() for initial graph
  - points() to add on additional data points

Resulting Plot

Pitch location by type
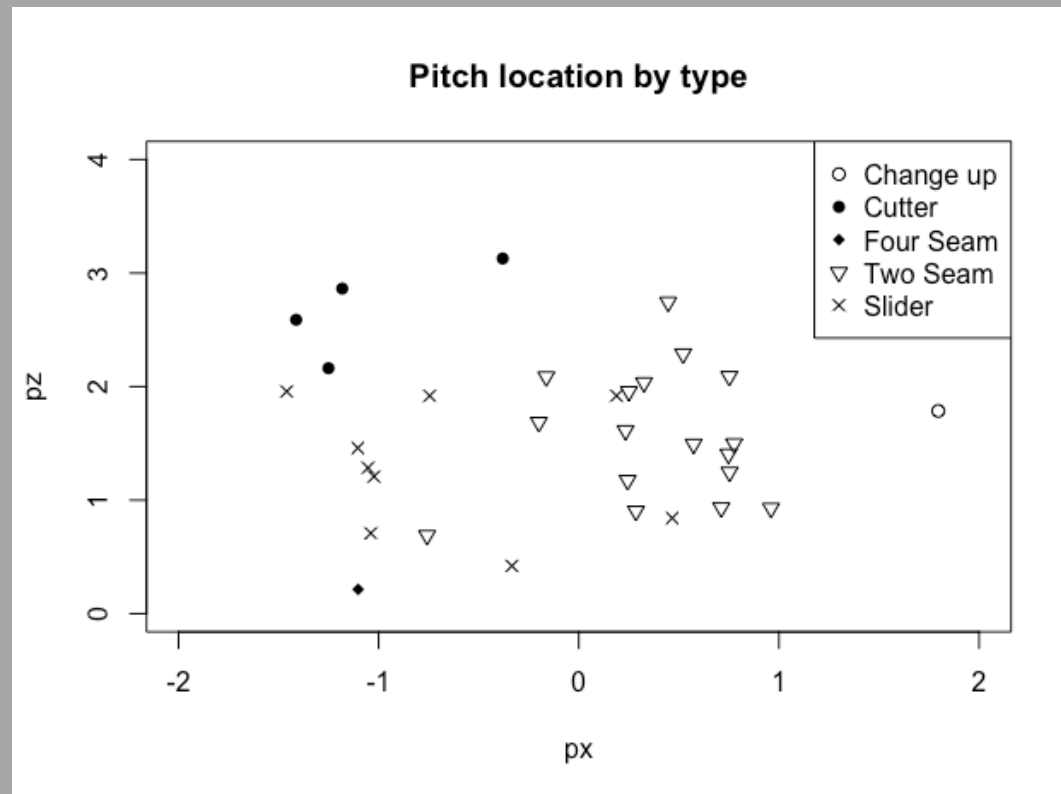
Resources

- Websites
  - Baseball Reference (www.baseball-reference.com)
  - Fan Graphs (www.fangraphs.com)
  - Society for American Baseball Research (www.sabr.org)
  - MIT Sloan Sports Analytics Conference (sloansportsconference.com)

- Books
  - Baseball Hacks by Joseph Adler
  - Analyzing Baseball Data with R by Max Marchi, Jim Albert
  - The Book by Tom Tango
  - Mathletics by Wayne Winston