Baseball: Hot and Cold Streaks

Malcolm Hess

March 16, 2015

Contributed by Malcolm Hess. This is based on my project investigating hot and cold streaks for hitters in baseball

Links

The code and files for this project can be found here: https://github.com/Mal-Hess/Baseball

If you enjoyed this or would like to see more of my work feel free to check out my github account for more projects. https://github.com/Mal-Hess/

Background

Baseball (America"s pastime) is a unique sport and the first that truly embraced statistical analysis. Baseball has by far the longest season in terms of games (162 games in a regular season) and also has one of the longest durations of play (from April to the end of September). The regular season for baseball is sort of like a marathon, one or two days won't make a big difference, what really matters is the player"s average by the end of the year.

Most hitters are known to go through "hot streaks" and "cold streaks" (also referred to as a slump). Players that are thought to have a cyclical nature in their hot and cold streaks but are expected to finish the year with a batting average close to what they had done previously. Some base ideas to support the 'streaks theory' is that players can be effected by a placebo effect from having success or failure. If a player is playing well, their confidence lends them to playing better but if a player is doing poorly, they lose confidence and their hitting suffers even further.

Another potential cycle in hitting success could be tied to the standard pitching rotation. Almost all major league teams start the season with their best pitcher (#1 pitcher) and try to have that player pitch in the most games possible. He would have

4 games of rest between starts. Which would mean that batters will face a “number 1” every 5th game. One would assume that a player to have consistently worse games against #1 pitchers.

This project will visualize the hitting success of an individual player throughout the season and hope to find visible proof of any kind of hitting cycle.

Objective

- Set up a system where I can easily create my own database of baseball data.
- Create a functions that can that will download the information and format it properly. Then save it to the local computer. This data should contain the career batting data for a specific player in a game by game bat log format.
- Create a function that will load all the data for a specific player.
- Create a visualization of hits across the year to investigate potential patterns in hot and cold streaks.

Implementation

Code Chunk 1: setting up the create a player function-

First I created the function which takes in three parameters. The player’s first name, last name, and key number. I also included a small amount of error-checking to make sure the values entered make sense. If there is no directory called Baseball this function will create it as a standard direcotry to store all of the baseball data.

```
createplayer <- function(playerfirstname, playerlastname, key=1){


  require(XML)


  require(RCurl)
```

```
#check if key is valid entry
if (class(key)!= "numeric"){
  stop("Invalid key: Requires number 1-9")
  stop}
if (key > 9 | key < 1){
  stop("Invalid key: Requires number 1-9")
  stop}
#keys are always two digit, so if less than 10 it makes
#key to a string and adds a zero to the front
if (key < 10){
  key <- substring(toString(key), 1, 1)
  key <- paste("0", key, sep="")
}
#checks to see if Baseball directory exists, and if not creates it.
```

```
  if(!file.exists("Baseball")){dir.create("/Baseball") }
  originalwd <- getwd()
  setwd("Baseball")
```

The next set of code cleans the user input and strings it together into the proper player identity. It also pulls from the website which career years the player has records in. It puts all those years into a vector and re-orders them from earliest to latest. I The the post-season is removed. I also included a check that makes sure the player doesn’t already exist in the local database.

```
playerfirstname <- as.character(playerfirstname)

  playerlastname <- as.character(playerlastname)

  #cleaning names and key to make player identity object

  subfirst <- substring(playerfirstname, 1, 2)

  sublast <- substring(playerlastname, 1, 5)

  identity <- paste(sublast, subfirst, key, sep="")

  identity <- tolower(identity)

  #checks to see if player already exists in local database

  filename <- paste0(identity, ".csv")
```

```r
if(file.exists(filename)){
  stop("Player already exists in database")
}
#making url to get to base page for the specified player
url <- paste0("http://www.baseball-reference.com/players/gl.cgi?id=", identity)
raw <- getURL(url)
data <- htmlParse(raw)
#making a list of all the years that this player has played in
xpath <- "//*[@id='stats_sub_index']/ul/li[4]/ul/li/a"
nodes <- getNodeSet(data, xpath)
years <- sapply(nodes, xmlValue)
#cleaning up the list of years, need to remove postseason and turn characters to numbers
years <- years[!is.element(years, "Postseason")]
years <- as.numeric(years)
```

```
 years<-sort(years)

 amountofyears <- length(years)
```

The next section gets one year worth of data and cleans it. The parameters are a player identity and a year. There were many obstacles I had to deal with in order transform the data into a clean and usable format. First I created the url and download the chart off of the website. I remove extra rows which were originally included to separate different months. I then created new columns which have the player identity and the year; this is used as an identifier column if multiple player's data is merged together. The dates have potential issues because 'double-headers' (a day where two games are played) have special symbols which I needed to remove in order to turn the dates into a proper date class object.

I fixed the "home/away" column which at first has no name, "H" is set for home games and "@" for away games. A new variable is made which is called "deltaavg", this is the change in batting average from one game to the next. I do not utilize this variable right now but I believe it could be interesting to use in future analysis.

```
getyeardata <- function(ident = identity, year=2014){

   #setting up URL to get data from a specific year

   url1<- "http://www.baseball-reference.com/players/gl.cgi?id="

   url2<- "&t=b&year="

   urlyear <- paste(url1, identity, url2, year, sep="")

   #downloading html site and taking out the table with the batting data

   html <- htmlTreeParse(urlyear, useInternal=TRUE)
```

```
tables <- readHTMLTable(html)
batlog<- tables$batting_gamelogs
rows<- nrow(batlog)
i<-1
while(i<=rows){ #removes Month rows
  if(batlog[i,1]=="April" | batlog[i,1]=="May" | batlog[i,1]=="June"|
       batlog[i,1]=="July"| batlog[i,1]=="August"| batlog[i,1]=="September"|
       batlog[i,1]=="October"){
    batlog <- batlog[-i,]
    i <- i - 1
    rows <- nrow(batlog)
  }
  i <- i + 1
}
```

```
    #adding a column to the front of the data that has identity and year on it

    #remove first column which is just row number, imported from html table.

    batlog <- batlog[,-1]

    batlog <- transform(batlog, Player=identity)

    batlog <- transform(batlog, Year=year)

    temp<- batlog[,36:37]

    batlog <- batlog[,-36:-37]

    batlog <- cbind(temp,batlog)

    #more data cleanup. Var.5 is currently the home/away column, away games signified with @

    #Date variable transformed to character so I can clean up the dates and later and a year to it.

    rows<- nrow(batlog)

    batlog <- transform(batlog, Var.5= as.character(Var.5), Date=as.character(Date), Gtm = as.character(Gtm))

    #double headers have extra symbols on them after the date, I need to

    #remove (1) or (2) from them to properly transform it to a date class.
```

```
#Gtm will also have extra () based on amount of games a player missed

#I am also adding the year to the end of the date.

i <-1

while(i<=rows){

  if(grepl("\\)", batlog$Date[i])) {

    nc <- nchar(batlog$Date[i])

    batlog$Date[i]<- substring(batlog$Date[i], 1, (nc-4))
  }

    batlog$Date[i]<- paste(batlog$Date[i], year, sep=", ")

  i <- i +1 }

batlog<- subset(batlog, Gcar != "Tm")

batlog<- subset(batlog, H != "HR")

#more data cleaning

colnames(batlog)[7] <- "Home"
```

```
  batlog <- transform(batlog, DELTAAVG= NA, BA = as.numeric(as.character(BA)), Home = as.character(Home))

  #One K loop to do two things.  First, deal with issues of home/away, second create deltaavg.

  #adds "H" (symbolize home game) to blank entires, away games are "@" symbol

  k<-1

  while(k<=nrow(batlog)){

    if(batlog[k,7]!= "@"){

      batlog[k,7] <- "H"

    }

    ##Makes new variable (deltaavg) that is the difference of Batting average from day to day

    if((k+1) <= nrow(batlog)){
      batlog$DELTAAVG[(k+1)]<-(batlog$BA[(k+1)] - batlog$BA[k])
    }
    k<-k+1
  } #end of k while loop
  batlog
}#end of getyeardata
```

I run through the entire vector and rbind all the clean years together. Lastly the directory is returned to where it was before the function began.

```
 #initializing object (careerdata) which will become the main dataframe

  careerdata <- NULL

  j<-1

  #Go through all years and rbind the data together into the careerdata object

  while (j <= amountofyears){

    b<- getyeardata(identity, year = years[j])

    careerdata <- rbind(careerdata, b)

    j<-j+1

  }

  filename <- paste0(identity, ".csv")

  write.csv(careerdata, file= filename)

  setwd(originalwd)

} #End of create player function!
```

This function is called loadplayer. It assumes that the createaplayer function has successfully been run on a specific player. In the future I will add an error catch in case the player does not exist in the local database. The player's career hitting data is returned as a dataframe.

```
loadplayer <- function (playerfirstname, playerlastname, key=1){

  #originalwd <- getwd()
  setwd("Baseball")

  if (key < 10){

    key <- substring(toString(key), 1, 1)

    key <- paste("0", key, sep="")

  }

  playerfirstname <- as.character(playerfirstname)

  playerlastname <- as.character(playerlastname)

  #cleaning names and key to make player identity object

  subfirst <- substring(playerfirstname, 1, 2)

  sublast <- substring(playerlastname, 1, 5)

  identity <- paste(sublast, subfirst, key, sep="")

  identity <- tolower(identity)
  filename <- paste0(identity, ".csv")

  dataframe <- read.csv(filename, header=TRUE)
  setwd("../")

  dataframe <- dataframe[,-1]

  dataframe

}
```

After having created the proper functions, I can run createplayer and then load player to get the career hitting logs for any baseball player I could want.

```
source('calendarheat.R')

createplayer("derek", "jeter", 1)
createplayer("josh", "hamilton", 3)
currentplayer <- loadplayer("derek", "jeter", 1)
hamilton <- loadplayer("derek", "jeter", 1)

simple <- transform(currentplayer, date = as.Date(Date, format = "%b %d, %Y"), h= as.numeric(H), Date=as.character(Date))
simple2 <- transform(hamilton, date = as.Date(Date, format = "%b %d, %Y"), h= as.numeric(H), Date=as.character(Date))
sub1 <- subset(simple, format(date, "%Y") %in% c("2014"))
sub2 <- subset(simple2, format(date, "%Y") %in% c("2013"))
```

Results

I made minor alterations to the original calendar heat code by changing the color pallet. The graph shows the number of hits on a given day. Dark blue is a bad day and red are good days.

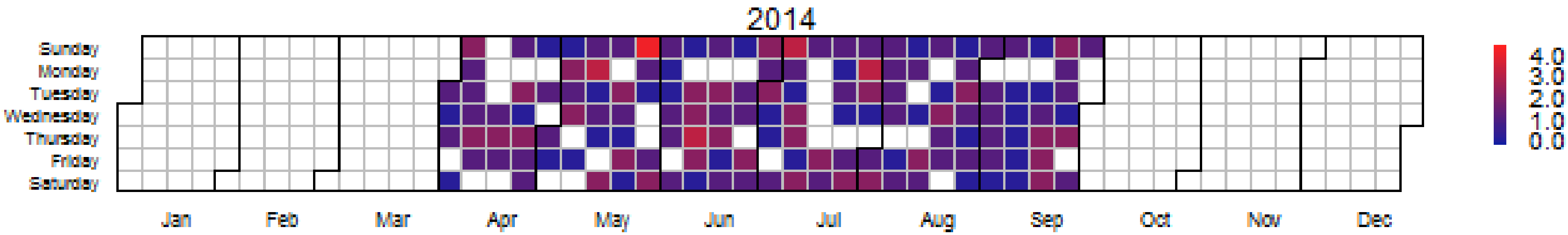
The chart is a new and interesting way to look at baseball data. It gives you a better feel for the strength of a player in a given time frame. After studying the calendar heat charts of dozens of players throughout multiple years of their careers I can say that there is no visible consistant cycle for success and failure for a specific player in a given year. I have also not found any evidence to support that certain players have a consistant and unique throughout multiple years of a career.

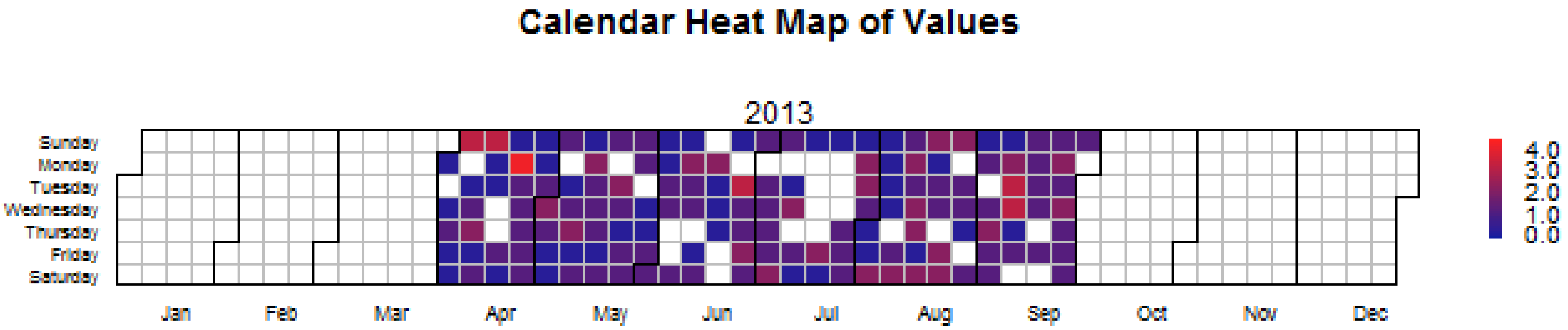
I think we can gleam very interesting realizations from the fact that these cycles do not exist. First of all there is no evidence to support that a player's success should cycle with the expected opponent's pitching rotation. We can't assume that players from year to year will be consistent in when they get their hits, whether it is consistent hitting throughout the year or hits clustered together in one or two months. Fans will still expect a player to have similar results as the previous year but how a player gets there seems to be incredibly random with no determinable cycle.

```
calendarHeat(sub1$date,sub1$h, date.form = "%b %d, %Y" )


calendarHeat(sub2$date,sub2$h, date.form = "%b %d, %Y" )
```

Calendar Heat Map of Values




Calendar Heat Map of Values
2013
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
4.0
3.0
2.0
1.0
0.0