IMT 573 Lab: Conditional Probability

Naga Soundari Balamurugan

October 30th, 2018

{Don't forget to list the full names of your collaborators!}

Collaborators:

Instructions:

1. Download the `week6a_lab.Rmd` file from Canvas. Open `week6a_lab.Rmd` in RStudio and supply your solutions to the assignment by editing `week6a_lab.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. If you are using more than just a standard function that you found from another source, please credit the source in the comments.
4. Collaboration on labs is encouraged, but students must turn in an individual assignments. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF` or `Knit Word`, rename the R Markdown file to `YourLastName_YourFirstName_Lab6a.Rmd`, knit a PDF or DOC and submit both the PDF/DOC and the Rmd file on Canvas.

In this lab, you will need access to the following R packages:

```
# Load some helpful libraries
library(tidyverse)
library(dplyr)
library(ggplot2)
```

Problem: If a baseball team scores X runs, what is the probability it will win the game?

This is the question we will explore in this lab (adapted from Decision Science News, 2014). We will use a dataset of baseball game statistics from 2010-2013.

Baseball is a game played between two teams who take turns batting and fielding. A run is scored when a player advances around the bases and returns to home plate. More information about the dataset can be found at http://www.retrosheet.org/.

Data files can be found on Canvas in the lab folder (look for 'cnames.txt' and all 'GL####.txt' files). Download the files and load them into one `data.frame` in R as shown below.

Comment this code to demonstrate you understand how it works.

```
#Read the csv file
colNames <- read.csv("cnames.txt", header=TRUE)
#Create an empty data frame
baseballData <- NULL
# Your comment here
for (year in seq(2010,2013,by=1)){
  # Your comment here
  mypath <- paste('GL',year,'.TXT',sep='')
  # Your comment here
  baseballData <- rbind(baseballData,read.csv(mypath,
  col.names=colNames$Name))
  baseballData <- tbl_df(baseballData)
}
```

Select the following relevant columns and create a new data frame to store the data you will use for your analysis.

- Date
- Home
- Visitor
- HomeLeague
- VisitorLeague
- HomeScore
- VisitorScore

```
#Create a list of the column names that are to be retained
variables_to_keep <-   c("Date", "Home", "Visitor", "HomeLeague", "VisitorLeague",
                         "HomeScore", "VisitorScore")

#Create a new dataframe with only the columns needed
baseBall_New <- baseballData[variables_to_keep]
```

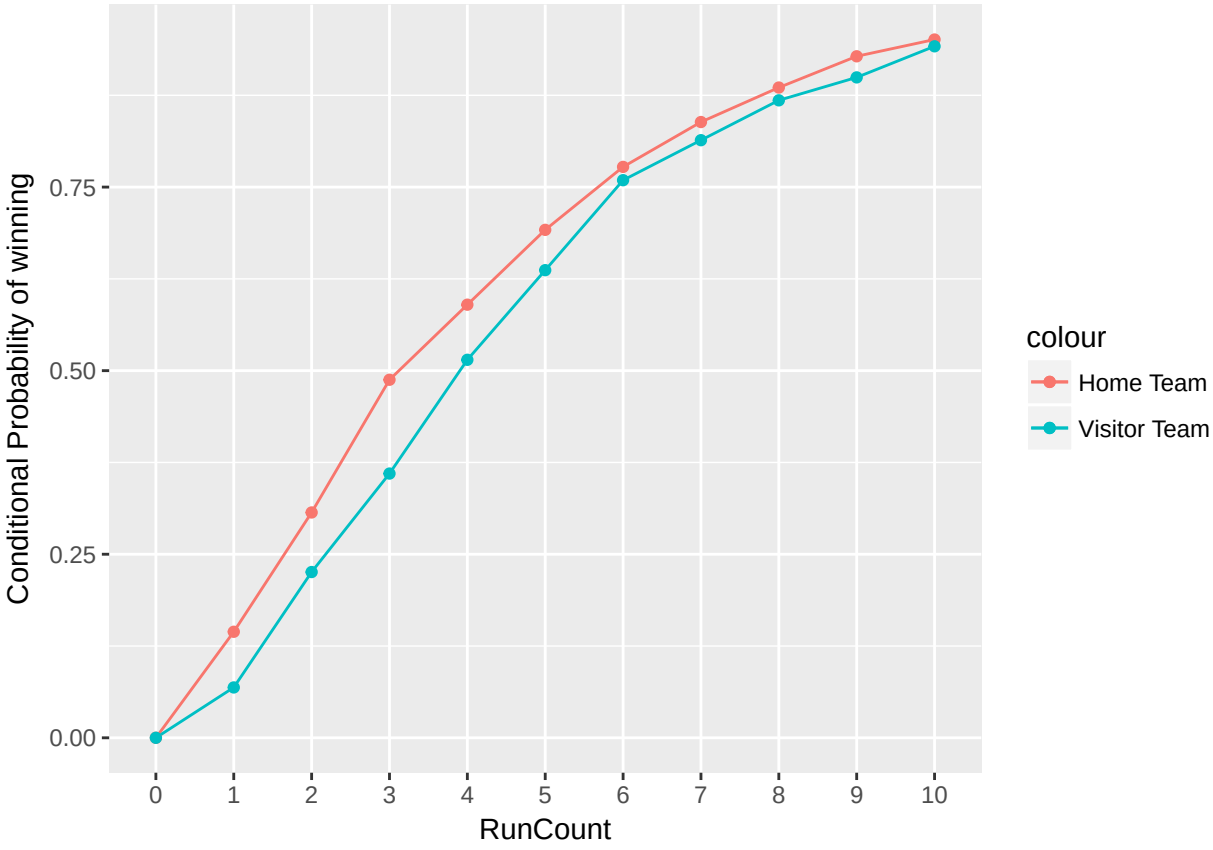
Considering only games between two teams in the National League, compute the conditional probability of the team winning, given X runs scored, for $X = 0, ..., 10$. Do this separately for Home and Visitor teams.

```
#Filter only the National Leagues
baseBall_NL <- baseBall_New %>% filter(HomeLeague == "NL")

#Add a new column that indicates the team won home team/visitor team
baseBall_NL <- baseBall_NL %>%  mutate(TeamWon = if_else(HomeScore > VisitorScore, 'H', 'V'))

#Computing number of times the home team and visitor team won
homeTeamWondf <- baseBall_NL %>% filter(TeamWon == 'H')
homeTeamWinCount <- nrow(homeTeamWondf)

visitorTeamWondf <- baseBall_NL %>% filter(TeamWon == 'V')
visitorTeamWinCount <- nrow(visitorTeamWondf)
```

```r
#Probability of winning for both teams
probHomeTeamWon <- homeTeamWinCount/nrow(baseBall_NL)
probVisitorTeamWon <- visitorTeamWinCount/nrow(baseBall_NL)

#Create dataframes that could store the conditional probabilities of hometeam/vistorteam winning
homeTeamdf <- data.frame(matrix(ncol = 2, nrow = 11))
colnames(homeTeamdf) <- c("RunCount", "HomeProbability")

visitorTeamdf <- data.frame(matrix(ncol = 2, nrow = 11))
colnames(visitorTeamdf) <- c("RunCount", "VisitorProbability")

condProbabilityHome <- function() {
  for(i in 0:10) {
    prob_i_runs <- nrow(baseBall_NL %>% filter(HomeScore == i))/nrow(baseBall_NL)
    prob_i_runs_Win <- nrow(homeTeamWondf %>% filter(HomeScore == i))/homeTeamWinCount
    prob_win_with_i_runs <- (prob_i_runs_Win * probHomeTeamWon)/prob_i_runs
    homeTeamdf$RunCount[i+1] <- i
    homeTeamdf$HomeProbability[i+1] <- prob_win_with_i_runs
  }
  homeTeamdf
}


condProbabilityVisitor <- function() {
  for(i in 0:10) {
    prob_i_runs <- nrow(baseBall_NL %>% filter(VisitorScore == i))/nrow(baseBall_NL)
    prob_i_runs_Win <- nrow(visitorTeamWondf %>% filter(VisitorScore == i))/visitorTeamWinCount
    prob_win_with_i_runs <- (prob_i_runs_Win * probVisitorTeamWon)/prob_i_runs
    visitorTeamdf$RunCount[i+1] <- i
    visitorTeamdf$VisitorProbability[i+1] <- prob_win_with_i_runs
  }
  visitorTeamdf
}


homeTeamdf <- condProbabilityHome()
visitorTeamdf <- condProbabilityVisitor()
conditionalProbDF <- merge(homeTeamdf, visitorTeamdf, by = "RunCount")
```

- Design a visualization that shows your results.

```r
ggplot(conditionalProbDF) +
geom_line(aes(x = factor(RunCount), y = HomeProbability, color = "Home Team", group = 1)) +
geom_line(aes(x = factor(RunCount), y = VisitorProbability, color = "Visitor Team", group = 2)) +
geom_point(aes(x = factor(RunCount), y = HomeProbability, color = "Home Team", group = 1)) +
geom_point(aes(x = factor(RunCount), y = VisitorProbability, color = "Visitor Team", group = 2)) +
  xlab("Runs Scored") + ylab("Conditional Probability of winning") +
  scale_x_discrete(name = 'RunCount')
```



- Discuss what you find.

  From the graph, we clearly know that the winning depends on the number of runs scored. For both the teams, the probability of winning is directly proportional to runs scored. Though the probability is slightly higher for the home team than the visitor team for the same runs scored.