INFX 573 Lab: Conditional Probability

Shuyang Wu

October 27th, 2016

Collaborators:

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

Instructions:

Before beginning this assignment, please ensure you have access to R and/or RStudio.

1. Download the `week5b_lab.Rmd` file from Canvas. Open `week5b_lab.Rmd` in RStudio (or your favorite editor) and supply your solutions to the assignment by editing `week5b_lab.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.
3. Be sure to include code chucks, figures and written explanations as necessary. Any collaborators must be listed on the top of your assignment. Any figures should be clearly labeled and appropriately referenced within the text.
4. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, rename the R Markdown file to `YourLastName_YourFirstName_lab5b.Rmd`, and knit it into a PDF. Submit the compiled PDF on Canvas.

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

```
# Load some helpful libraries
library(tidyverse)
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 (ddapted from Decision Science News, 2014). We will use a dataset of baseball game statistics from 2010-2013.

Baseball is a 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/`.

Note: More information about the dataset can be found at http://www.retrosheet.org/

```
colNames <- read.csv("cnames.txt", header = TRUE)  #read cnames into colName
baseballData <- NULL  #create empty dataframe
for (year in seq(2010, 2013, by = 1)) {
    mypath <- paste("GL", year, ".TXT", sep = "")  #find the paths of files containing 'GL' + year from 20
    # cat(mypath,'\n')
    baseballData <- rbind(baseballData, read.csv(mypath,
        col.names = colNames$Name))  #bind data from each file into the dataframe
    baseballData <- tbl_df(baseballData)  #forward the argument to as.data.frame
}
# baseballData
myvars <- c("Date", "Home", "Visitor", "HomeLeague",
    "VisitorLeague", "HomeScore", "VisitorScore")
relev_baseballData <- baseballData[myvars]
```

Data files can be found on Canvas in the lab folder. 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.*

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

```
myvars <- c("Date", "Home", "Visitor", "HomeLeague",
    "VisitorLeague", "HomeScore", "VisitorScore")
relev_baseballData <- baseballData[myvars]
```

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.*

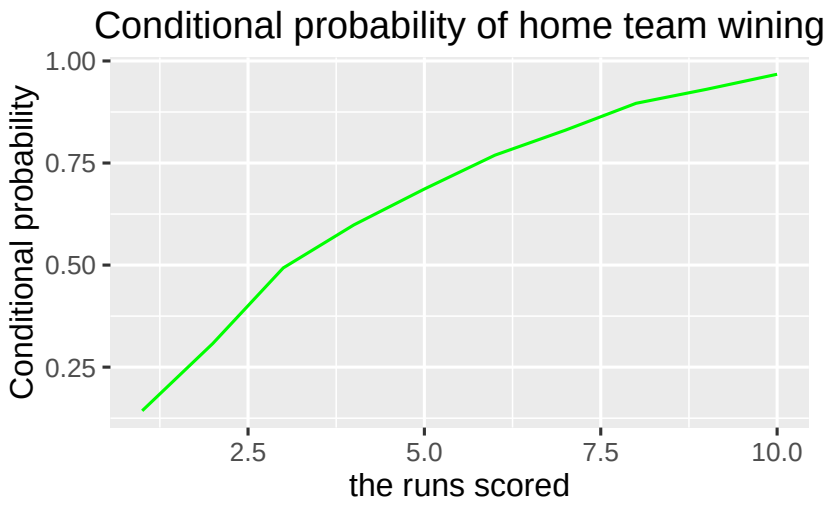
- Design a visualization that shows your results.
- Discuss what you find.

```
# select only national leagues
cp <- subset(relev_baseballData, HomeLeague ==
```

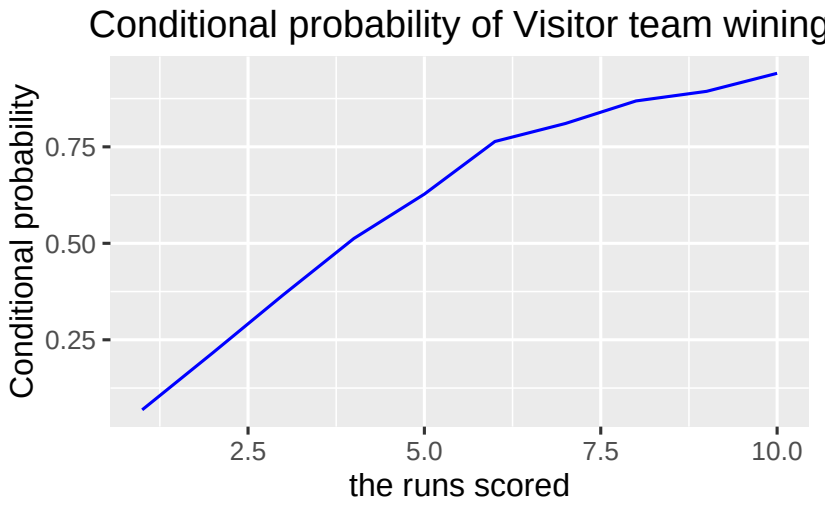
```
    "NL" & VisitorLeague == "NL")
# create columns to mark winning team
if (cp$HomeScore > cp$VisitorScore) {
    cp$HomeWin <- TRUE
}
```

```
## Warning in if (cp$HomeScore > cp
## $VisitorScore) {: the condition has length >
## 1 and only the first element will be used
```

```
cp$HomeWin <- ifelse(cp$HomeScore > cp$VisitorScore,
    1, ifelse(cp$HomeScore < cp$VisitorScore,
        0, NA))
cp$VisitorWin <- ifelse(cp$HomeScore < cp$VisitorScore,
    1, ifelse(cp$HomeScore > cp$VisitorScore,
        0, NA))
cp$HomeWin <- as.logical(cp$HomeWin)
cp$VisitorWin <- as.logical(cp$VisitorWin)
# find conditional probabilities for each
# number of runs scored and visualization
homewinprop <- cp %>% group_by(HomeScore) %>%
    summarise(prop_win = (prop.table(table(HomeWin)))[2])
homewinprop <- homewinprop[which(homewinprop$HomeScore >=
    1 & homewinprop$HomeScore <= 10), ]
ggplot(homewinprop, aes(HomeScore, prop_win)) +
    geom_line(color = "green") + ggtitle("Conditional probability of home team wining") +
    xlab("the runs scored") + ylab("Conditional probability")
```



```
visitorwinprop <- cp %>% group_by(VisitorScore) %>%
    summarise(prop_win = (prop.table(table(VisitorWin)))[2])
visitorwinprop <- visitorwinprop[which(visitorwinprop$VisitorScore >=
    1 & visitorwinprop$VisitorScore <= 10), ]
ggplot(visitorwinprop, aes(VisitorScore, prop_win)) +
    geom_line(color = "blue") + ggtitle("Conditional probability of Visitor team wining") +
    xlab("the runs scored") + ylab("Conditional probability")
```



The probability of winning increases as more runs being scored.

Extra Credit: Repeat the above problem, but now consider the probability of winning given the number of hits.