Progress Report for April 4th, 2016

Tom Jeon & Steele Valenzuela

April 1, 2016

Introduction

We would like to “borrow” an idea from Andrew Gelman’s blog pertaining to sports. Initially, when sports fanatics and statistics mesh, betting and gambling on who will win a game becomes the most salient information they care about. That was our idea, too, because everyone wants to predict a winner. What is not as exciting, as enlighted by Andrew Gelman (http://andrewgelman.com/2014/02/25/basketball-stats-dont-model-probability-win-model-expected-score-differential/) is to model the expected score differential.

Score differential, which is the difference of the final score, of which requires two teams to be matched up against one another. We have chosen the National Basketball Association (NBA) as our sport of interest. There are **30 teams**, each playing **82** games, which is a moderate amount when one considers the number of games played by professional football teams (not as many games) and baseball teams (almost twice the amount of games). This presents us with a manageable number of games, or for this case, observations.

Moreover, we will attempt to fit a Bayesian hierarchical model as well as adding a time-series extension to the model, something which we have not yet explored in class. For the hierarchical model, the set-up includes observing the score-differential for one team, the Boston Celtics, which is nested within a division along with a few teams, and where the division is nested within a larger conference, which is composed of half of all teams. Therefore, we will create a multi-level model that will draw information from the divisions, conferences, and finally, the time-series aspect will draw from prior games.

Data

The following code displays how we extracted `HTML` tables from a website to create our data sets as well as a few steps in cleaning the data set.

```r
url <- read_html("http://www.basketball-reference.com/leagues/NBA_2015_games.html")
tbl <- html_table(url)

d1 <- tbl_df(as.data.frame(tbl[1]))  # data set we'll be using for this report
d2 <- tbl_df(as.data.frame(tbl[2]))  # playoffs. not using this one unless we get far in the project

## Cleaning
d1$Visitor.Neutral <- as.factor(d1$Visitor.Neutral)
d1$Home.Neutral <- as.factor(d1$Home.Neutral)

names(d1) <- c("Date", "ET.Start", "Var.3", "VisitingTeam", "PTS.V", "HomeTeam",
    "PTS.H", "Var.8", "Notes")
# names(d1) #str(d1)
home <- d1 %>% filter(HomeTeam == "Boston Celtics")
visiting <- d1 %>% filter(VisitingTeam == "Boston Celtics")

home$score.dif <- home$PTS.H - home$PTS.V  #negative values denote celtics' loss
visiting$score.dif <- visiting$PTS.V - visiting$PTS.H
kable(head(home))  # display table
```

Date	ET.Start	Var.3	VisitingTeam	PTS.V	HomeTeam	PTS.H	Var.8	Notes
Wed, Oct 29, 2014	7:30 pm	Box Score	Brooklyn Nets	105	Boston Celtics	121		
Wed, Nov 5, 2014	7:30 pm	Box Score	Toronto Raptors	110	Boston Celtics	107		
Fri, Nov 7, 2014	7:30 pm	Box Score	Indiana Pacers	98	Boston Celtics	101		
Wed, Nov 12, 2014	7:30 pm	Box Score	Oklahoma City Thunder	109	Boston Celtics	94		
Fri, Nov 14, 2014	7:30 pm	Box Score	Cleveland Cavaliers	122	Boston Celtics	121		
Mon, Nov 17, 2014	7:30 pm	Box Score	Phoenix Suns	118	Boston Celtics	114		

We see the first 6 rows of the *home* data set, which displays the results for the Boston Celtics, the team they played, and the score differential. The *visiting* data set may also be viewed.

Team Conferences and Divisions

In the NBA, there are two conferences; Eastern & Western. And in each conference, there are 3 divisions. For the Eastern conference, the divisions are; Atlantic, Central, Southeast. For the Western conference, the divisions are; Northwest, Pacific, Southwest. These may be verified from the website where pulled the original data which displayed the 2014-15 schedule (http://www.basketball-reference.com/leagues/NBA_2015_standings.html).

Now, let's link each team (from the *home* and *visiting* dataframes) to their respective conference and division, which will make up two levels for our hierarchical model.

```
atl.div <- c("Toronto Raptors", "Boston Celtics", "Brooklyn Nets", "Philadelphia 76ers", "New York Knic
cen.div <- c("Cleveland Cavaliers", "Chicago Bulls", "Milwaukee Bucks", "Indiana Pacers", "Detroit Pist
se.div <- c("Atlanta Hawks", "Washington Wizards", "Miami Heat", "Charlotte Hornets", "Orlando Magic")
nw.div <- c("Portland Trail Blazers", "Oklahoma City Thunder", "Utah Jazz", "Denver Nuggets")
pac.div <- c("Golden State Warriors", "Los Angeles Clippers", "Phoenix Suns", "Sacramento Kings", "Los
sw.div <- c("Houston Rockets", "San Antonio Spurs", "Memphis Grizzlies", "Dallas Mavericks", "New Orlea
east.conf <- c(atl.div, cen.div, se.div)
west.conf <- c(nw.div, pac.div, sw.div)
```

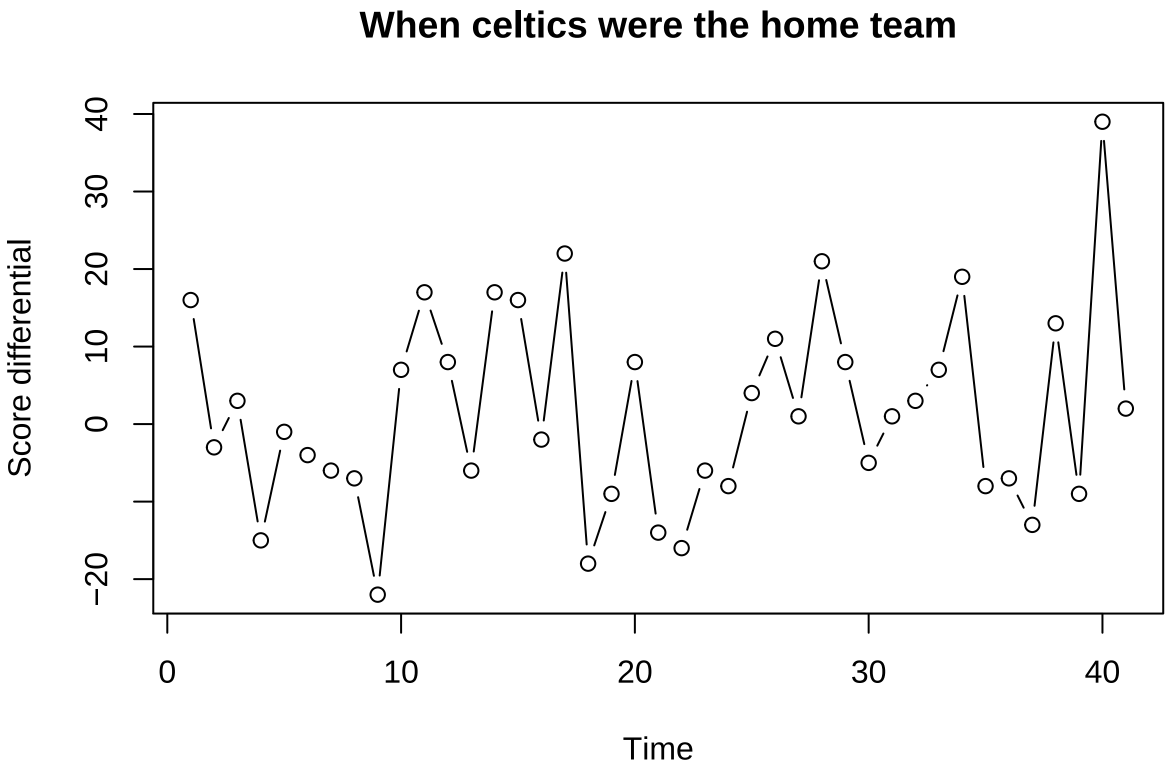
And let's add the division and conference grouping variables to both the *home* and *visitor* data frames.

```
home$vtconf <- ifelse(home$VisitingTeam %in% east.conf, "East", "West") # visiting team conference
home$htconf <- "East"
home$vtdiv <- ifelse(home$VisitingTeam %in% atl.div, "Atlantic Division",
                     ifelse(home$VisitingTeam %in% cen.div, "Central Division",
                            ifelse(home$VisitingTeam %in% se.div, "Southeast Division",
                                   ifelse(home$VisitingTeam %in% nw.div, "Northwest Division",
                                          ifelse(home$VisitingTeam %in% pac.div, "Pacific Division", "S
home$htdiv <- "Atlantic Division"

visiting$htconf <- ifelse(visiting$HomeTeam %in% east.conf, "East", "West")
visiting$vtconf <- "East"
visiting$htdiv <- ifelse(home$VisitingTeam %in% atl.div, "Atlantic Division",
                     ifelse(home$VisitingTeam %in% cen.div, "Central Division",
                            ifelse(home$VisitingTeam %in% se.div, "Southeast Division",
                                   ifelse(home$VisitingTeam %in% nw.div, "Northwest Division",
                                          ifelse(home$VisitingTeam %in% pac.div, "Pacific Division", "S
visiting$vtdiv <- "Atlantic Division"
```

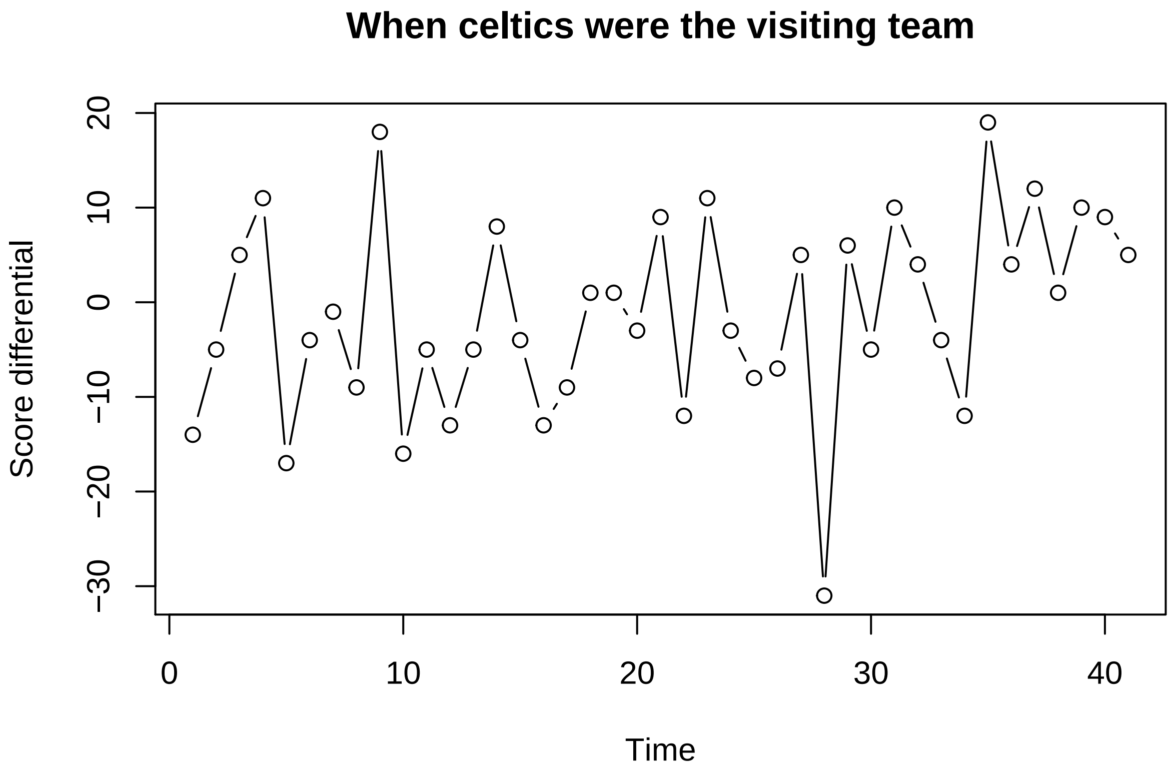
Time Series

Next, let's view the *score differential* on the *y-axis* and *time* on the *x-axis* for all home and away games.

```
plot(home$score.dif, type = "b", xlab = "Time", ylab = "Score differential", main = "When celtics were t
```



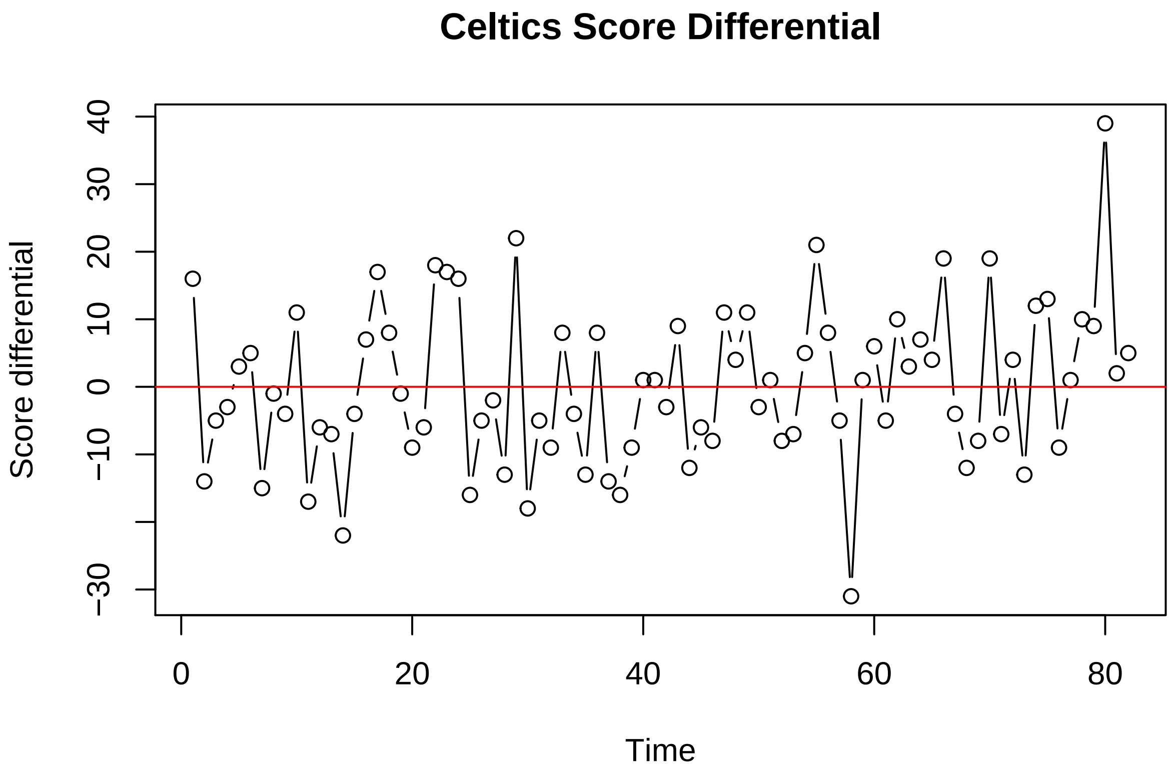
```
plot(visiting$score.dif, type = "b", xlab = "Time", ylab = "Score differential", main = "When celtics w
```



Next, let's view the progression of the Boston Celtics over the whole season, disregarding *home* and *away*.

```
## CLASSIC TIME SERIES

#Combine the home and visiting by proper time

#Have to convert Date column to date data type
#How to use lubridate?
home <- home[,-c(2,3,8,9)]
visiting <- visiting[, -c(2,3,8,9)]

dates <- as.Date(home$Date, "%a, %b %d, %Y")
home$Date <- dates

dates2 <- as.Date(visiting$Date, "%a, %b %d, %Y")
visiting$Date <- dates2
#View(home)

games <- rbind(home, visiting)
games <- games %>% arrange(Date)
#View(games)

#convert VisitingTeam == "Boston Celtics" to 1, 0 otherwise
#convert HomeTeam == "Boston Celtics" to 1, -1 otherwise
#Final result: VisitingTeam + HomeTeam = 1 if Home, 0 otherwise
games$VisitingTeam <- ifelse(games$VisitingTeam == "Boston Celtics", 1, 0)
games$HomeTeam <- ifelse(games$HomeTeam == "Boston Celtics", 1, -1)
games$Home <- games$VisitingTeam + games$HomeTeam
#View(games)
games <- games[, -c(2,3,4,5)]
```

```
#View(games)

plot(games$score.dif, type = "b", main = "Celtics Score Differential", xlab = "Time", ylab = "Score dif
abline(h = 0, col = "red")
```



```
#Further steps:
## time series model specification
## hierarchy based on teams, div, conf, home/away
```

The final plot shows time-series data for the Boston Celtics.

Next Steps

Further steps would include specifying the most appropriate autoregressive integrated moving average (ARIMA) model and its model diagnostics, **finding and merging data from other sources** (levels complete, so covariates?), fitting a Bayesian hierarchical level model, specifying those levels, and merging possible covariates.

Questions (for project members)

1. If we combine both home and away data sets, will that be another level?
2. Can you provide more info on ARIMA? I've heard of moving averages, but only briefly, and in the context of splines.
3. How will covariates play into time-series? Are they updated after the passing of one-unit of time? In our case, after one game.
4. Not a question, but let's find a time-series example within a Bayesian context.