Sports Analytics Using Bayesian Methods

Tom Jeon and Steele Valenzuela

April 26, 2016

Contents

Introduction:

The recent explosion of Fantasy Sports engendered many sports fans to rely on hard truths based on real data. When sports fanatics and statistics mesh, betting and gambling on who will win a game become the most salient information. The most common way sports analysts go about this is to train models that directly estimate the probability of a win; however, considering very close games or blowouts, there's a lot of information in the score differential that is disregarded if only the outcome of the game is modeled. In this report, we use Bayesian methods to model expected score differential for the Boston Celtics, using data from 2004 to 2016. We explore time series and spline regression in the Bayesian framework to answer the following questions. Who is a better head coach, Doc Rivers, the former head coach of the Boston Celtics from 2004 to 2013, or Brad Stevens, the current head coach? What is the expected score differential for the next Celtics game? How does the expected score differential change if given the opposing team's ranking?

Data:

The data we collected from sports-reference.com included the final score records of all games from 2004 to 2016, and team rankings for each corresponding year. We subsetted for when the Boston Celtics were either the Home Team or the Visiting Team and then calculated the score differential for each game, then assigned each observation the ranking of the opposing team for each season.

Date	diff	Team	Rk
2003-10-29	23	Miami Heat	26
2003-10-31	2	Memphis Grizzlies	22
2003-11-01	-7	New Orleans Hornets	14
2003-11-05	-8	Detroit Pistons	5
2003-11-07	-7	New Jersey Nets	4

Because many separate attributes about the team and its players were aggregated to calculate the rankings of each team, we decided the data we had were sufficient. The opposing teams' rankings for each season were based on the previous season's data.

Scraping, Cleaning, and Management

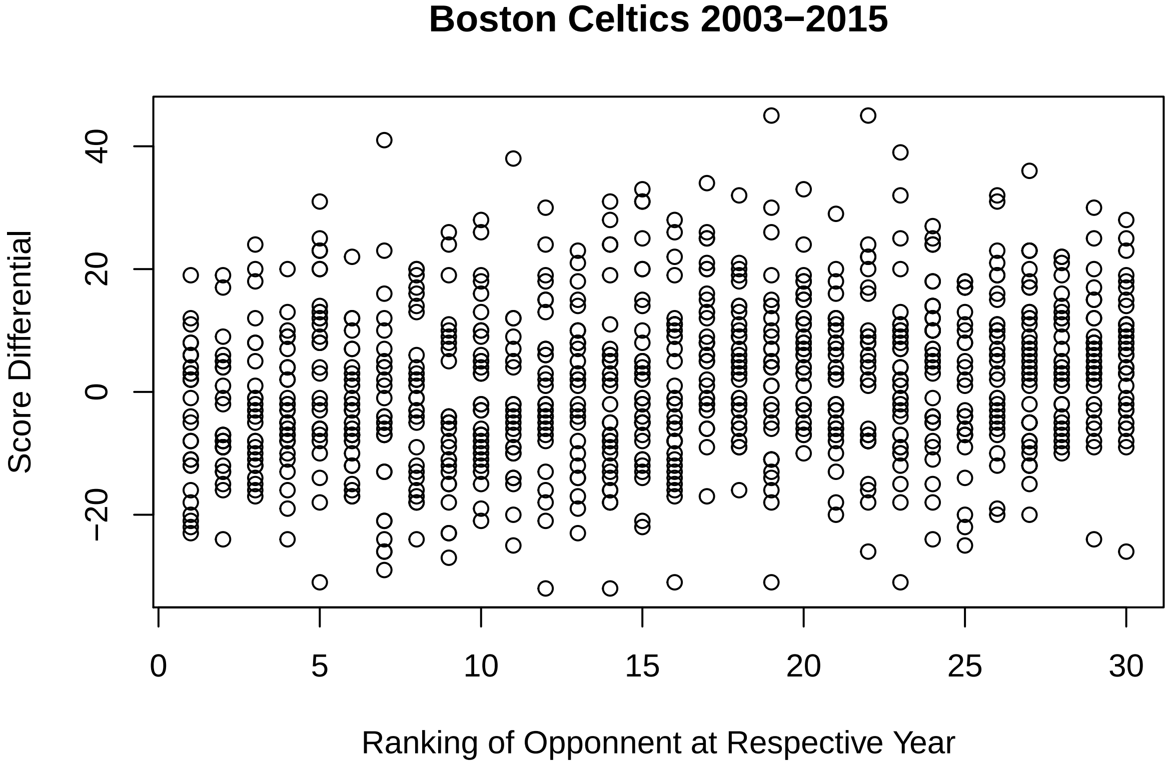
One would think that the most popular place to collect data is from ESPN. Unfortunately, the site is not user-friendly for those looking to do their own analyses, but rather for those who simply want to view the box-scores. Luckily, we were familiar with other sports web sites and chose to scrape data from the site [sports-reference.com][sports-reference.com].

In the most simple of terms, we needed data that identified a game, the two teams that played this game, and their respective scores. From there, we could calculate our response variable of interest, *score differential*, which is calculated by taking the difference of scores (*Please see appendix for a screenshot of the website*). After using the `rvest` library and two of its functions, we're able to extract the data and place it into a suitable dataframe that we can work with in `R`.

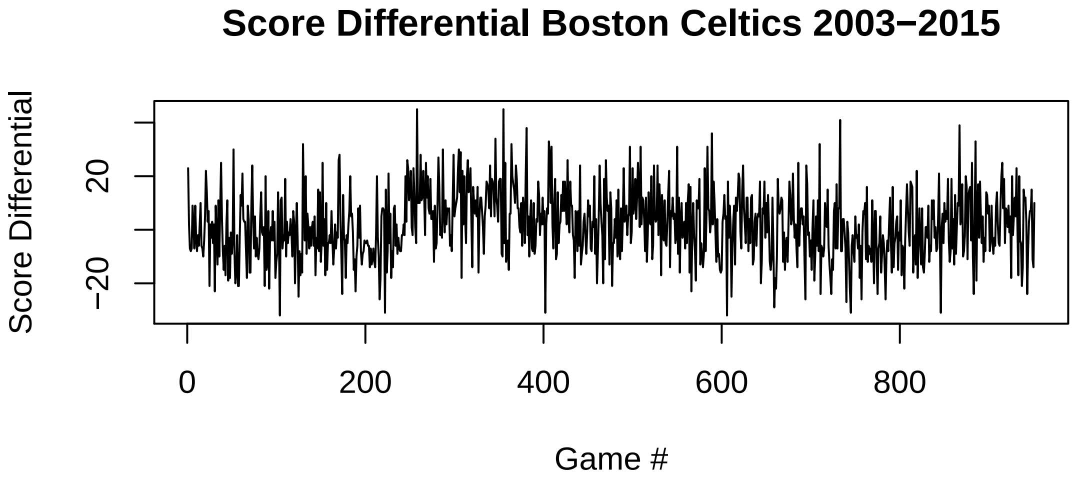
The libraries `dplyr` and `tidyr` were also crucial in cleaning the data by renaming variables, deleting columns, switching data from a wide to long format, adding attributes to each team, and lastly, filtering data by team, which we did throughout this project in order to focus on the local team, the Boston Celtics.

Exploratory Analysis

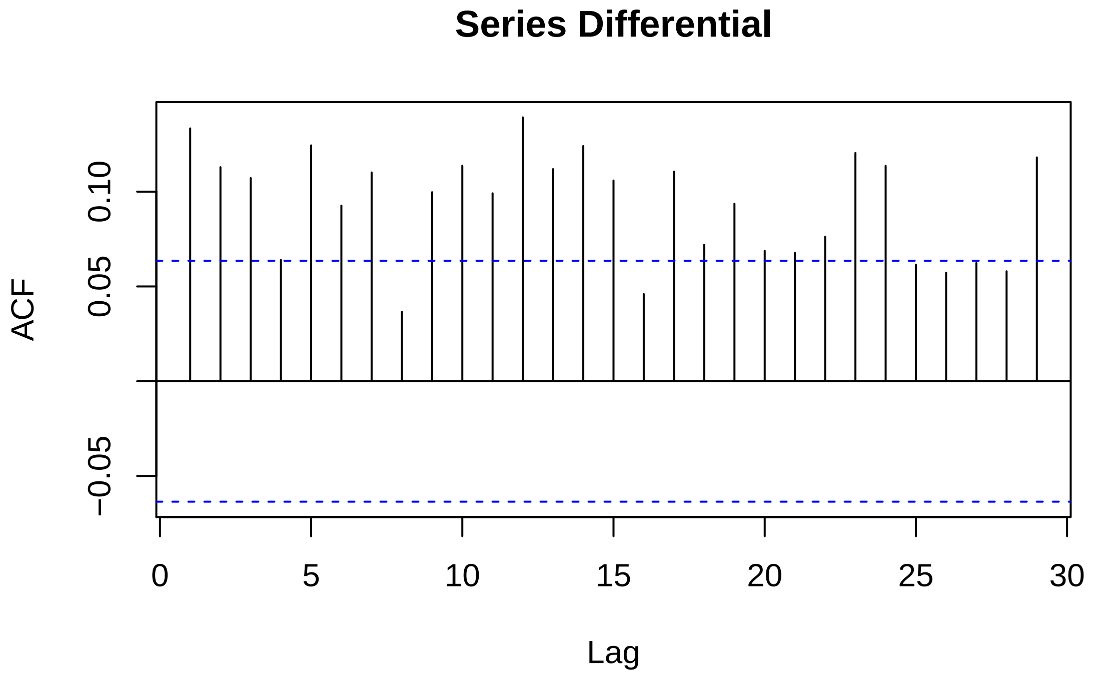
The first graph shows the score differential between the Celtics and their opponent at the end of every game from 2004 to 2015. This means that because there were multiple games between the Celtics and the 1 to 30 ranked teams, there are many y values per x value. Here, we defined a positive score differential to mean that the Celtics had a higher score at the end of the game and won.



Next we looked the score differentials as a function of time to explore whether it was appropriate to use an AR(1) model for the score differentials.



The data looks stationary, with no obvious heteroskedastic qualities. To see if there were any correlation between the current and previous games' score differentials, we plotted an autocorrelation function for this data.



Looks like there are significant correlations between previous games and the current one! An AR(1) process to model the score differentials seems appropriate for our data.

Methods & Results

Doc Rivers vs. Brad Stevens

One question we can answer using this data is, who is a better head coach, Doc Rivers or Brad Stevens? We approached this question by fitting two Bayesian B-splines models, one with the data for when Doc Rivers was head coach, and one for Brad Stevens.

Here is the model set up:

$$y_i \sim N(\mu_i, \sigma_i^2)$$

$$\mu_i \sim \sum_{k=1}^{K} (b_k(x_i))$$

With priors:

$$\alpha_k \sim N(0, 100)$$

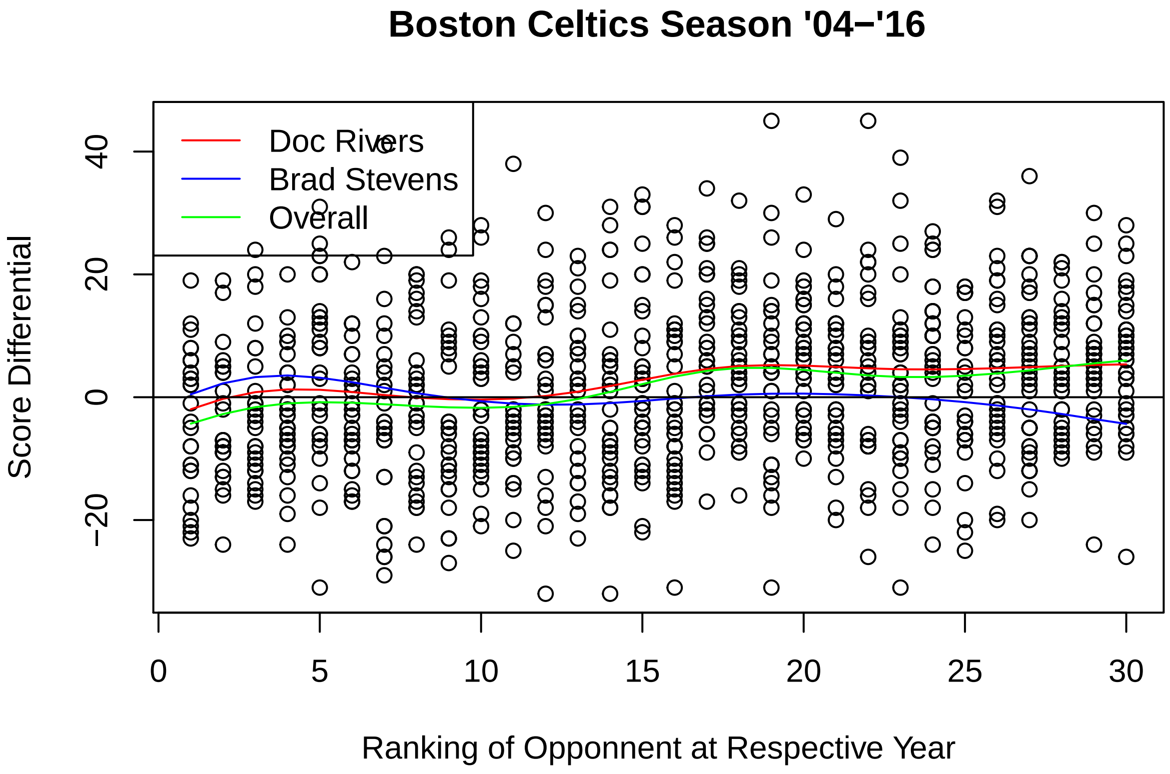
$$\sigma_y \sim U(0, 3)$$

Spline intervals and degree:

$$I = 2.5$$

$$degree = 3$$

To easily compare our splines models, we plotted both models on the same graph as shown below.



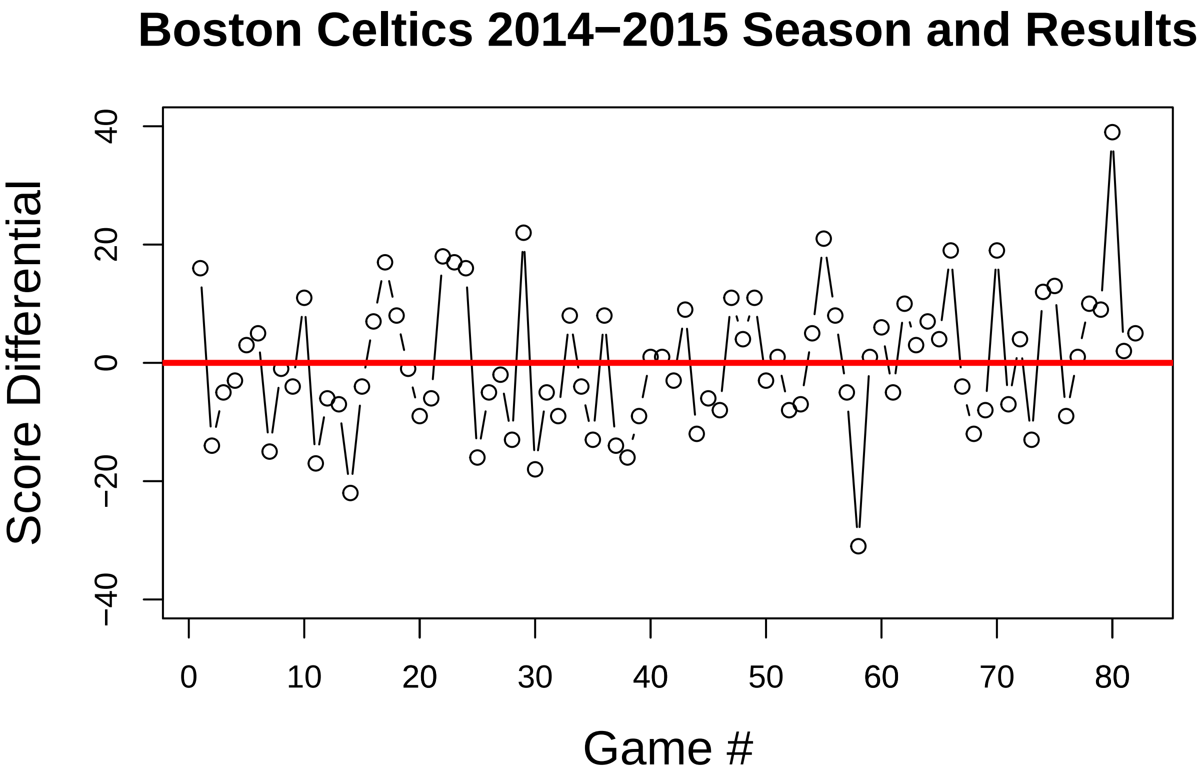
As expected, the green spline model which represents the whole data regardless of whether the Boston Celtics were coached under Doc Rivers or Brad Stevens shows there is an upward trend as the ranking of the opposing team increases. That is, the Celtics seem to do better against teams that are ranked lower. A similar trend follows the red spline model, which represents Celtics performance under Doc Rivers, except it is slightly above the green one, which suggests that under Doc Rivers, the Celtics performed better than how the Celtics performed with and without Doc Rivers. Now if you look at the blue line, the spline model that represents Celtics' performance under the current head coach, Brad Stevens, there is a downward trend! This suggests that under Brad Stevens, the Celitcs do better against teams that are ranked higher. We thought this revealed a lot about Stevens' coaching style.

To have a quantitative way of comparing the two head coaches, we looked at the area under the splines curves for Rivers and Stevens (red and blue) and over the horizontal line at $y = 0$. Once we do so, we can effectively conclude that the Celtics performed better under Doc Rivers and hence Rivers is the superior coach, holding all other variables constant.

Time Series Comparison of the Frequentist and Bayesian AR(1) Process

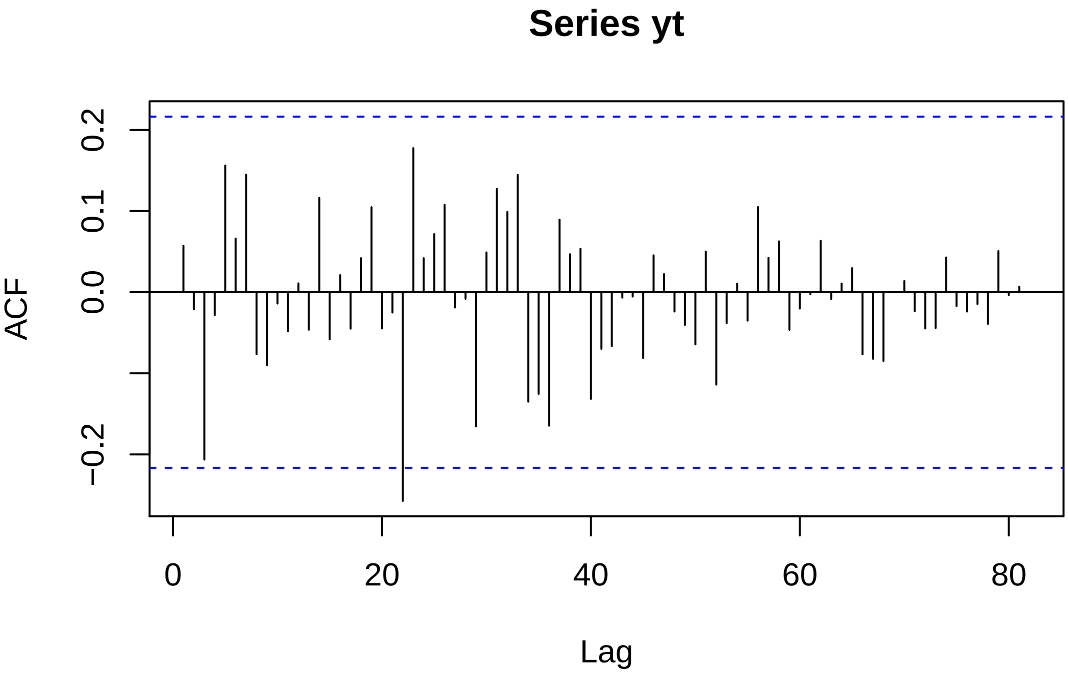
2014 - 2015 Season

For our initial analysis, team member Steele Valenzuela initially explored and exhausted exploratory analyses on how to fit an AR(1) process on observed data, particulary *score differential*, which was the y_t. For homework 7, a Bayesian AR(1) process was utilized for simulated data, but not observed data, which is why this initial process was undertaken and then later applied to the compiled dataset featuring additional games from the 2003-04 to 2014-15 seasons. Below is a time series plot of the Boston Celtics 2003-2004 season with *score differential* as y_t, observed at each of the 82 games.



Can we infer anything from this plot? Maybe not as it does appear to be a sequence of random variables over 82 games.

Let’s investigate further with a plot of the auto-correlation function.



The correlation between one game, between two games, and so on, does not present a clear pattern.

Next, let's estimate σ and ρ. If we use the built-in functions in `R`, we can create a lag varible, which is then fit in a model against our original response variable y_t, which outputs ρ. Additionally, ρ and σ can be obtained from the `arima(...)` function as seen below.

```
ylag1 <- lag(y, 1); y1 <- cbind(y, ylag1)
(ar1fit <- lm(as.numeric(y1[, 1]) ~ as.numeric(y1[, 2])))
```

```
##
## Call:
## lm(formula = as.numeric(y1[, 1]) ~ as.numeric(y1[, 2]))
##
## Coefficients:
##         (Intercept)  as.numeric(y1[, 2])
##            -0.04269              0.05728
```

```
arima(x = yt, order = c(1, 0, 0))
```

```
##
## Call:
## arima(x = yt, order = c(1, 0, 0))
##
## Coefficients:
##         ar1  intercept
##       0.058      0.174
## s.e.  0.111      1.363
##
## sigma^2 estimated as 135.3:  log likelihood = -317.57,  aic = 639.13
```

For both outputs, we see that $\rho = 0.058$. For the last output, $\sigma = \sqrt{135.3} = 11.63$. σ and ρ are found by frequentist methods, specifically maximum likelihood or minimize conditional sum-of-squares, which is beyond the scope of this class.[*citation*]

Now, onto a comparison of the Bayesian AR(1) process. Let us propose the following model:

$$y_t \sim AR(1)$$

$$\rho \sim U(-1, 1)$$

$$\sigma \sim U(0, 15)$$

By constraining ρ between -1 and 1, we are assuming a stationary process for y_t. In the frequentist method, σ was estimated to be 11.63, hence the upper bound is set to be larger with a value of 15. In order to begin the process, we must estimate y_1. Using the stationary distribution, $y_1 \sim N(0, \sigma^2/(1-\rho^2))$, which follows in setting up the remainder of the y_t's. Our results for the model fit are as follows:

	freq.param	mean	sd	Lower.Bound	Upper.Bound	freq.Lower.Bound	freq.Upper.Bound
sigma	11.632	11.895	0.962	10.181	13.960	NA	NA
rho	0.058	0.059	0.114	-0.173	0.282	-0.15956	0.27556

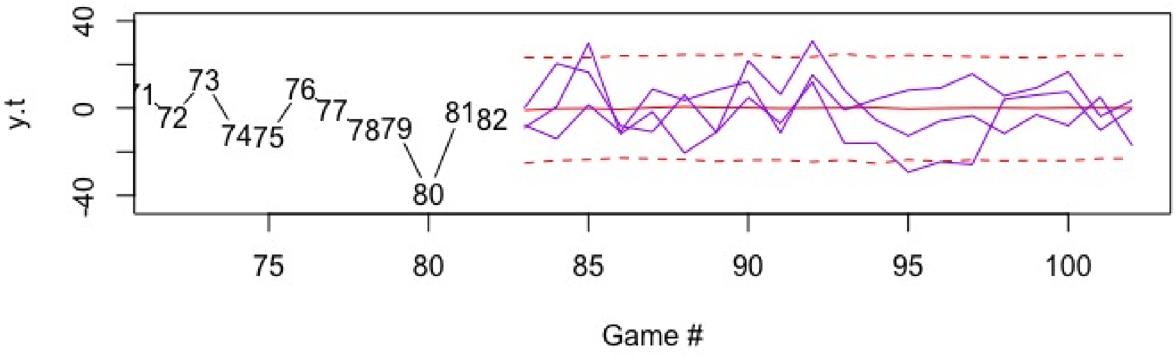
For convergence diagnostics on both σ and ρ, traceplots checked out, N_{eff} was equivalent to 4,000 and 1500, respectively, and the $\hat{R}$ was close to 1. As for the model fit, we see that the Bayesian methods produced estimates close to the frequentist estimates as the mean for σ and ρ are relatively close. We see for the

Bayesian AR(1) process, the 95% credible interval is wider than the frequentist interval. Not shown here, but the reported median was 11.812 and 0.061 for σ and ρ.

Lastly, to mirror the notes on Time Series for observed data, forecasted trajectories were created. From the slides, we see that given a posterior sample $\rho^{(s)}$ and $\sigma^{(s)}$, one forecast trajectory $y_{t+p}^{(s)}$ for $p \geq 1$ can be obtained as follows:

$$y_{t+p}^{(s)}|\rho^{(s)}, \sigma^{(s)} \sim N(\rho^{(s)} y_{t+p-1}^{(s)}, \sigma^{(s)2})$$

where $y_t^{(s)} = y_t$ (observed). In this particular case, $t = 82$, or the 82nd game, and $p = 20$, or a trajectory of 20 games up to 102 games. Here is the following forecast:



Up to game 82, we see the observed score differential, the red solid line is the point forecast, which is close to zero, the red-dotted lines are the 95% predictive interval for the point forecast, and the purple lines indicate example trajectores, which we see cross the bounds of the predictive intervals and is just all over the place, as expected.

Compiled Season of the Boston Celtics

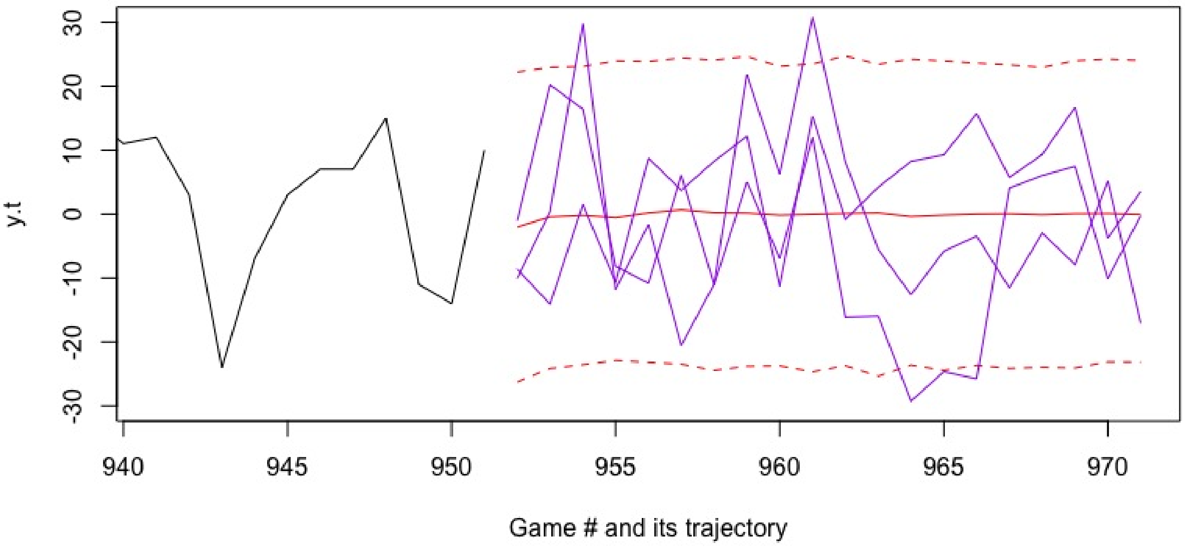
Next, let's view a larger data set, which includes the score differential for the Boston Celtics from seasons 2003-2004 to 2014-2015, 951 games. Although we could have only included a fully-fitted model, we first worked with data that involved a single season due to the meticulousness of collecting multiple seasons, but it was completed.

Similar to the previous section with only the 2014 - 2015 season, our parameters σ and ρ will utilize the same prior distributions. Convergence and traceplots have checked out, so there is no need to report any concerns. Here are the results:

	freq.param	mean	sd	Lower.Bound	Upper.Bound	freq.Lower.Bound	freq.Upper.Bound
sigma	12.566	12.677	0.299	12.117	13.273	NA	NA
rho	0.134	0.150	0.032	0.088	0.213	0.07128	0.19672

Again, similar analyses, but we can dissect this. First, we notice that that the frequentist and Bayesian ρ differ by 0.016. This may or may not seem like a significant difference, but one would think that as more data was collected, the estimates would line up closer to one another, like the analyses for the 2014 - 2015 season. Additionally, we see the 95% credible intervals differ from the 95% confidence intervals, which may be due to the shift in ρ estimates.

Lastly, let us add the forecasted trajectory:



Again, we see sampled trajectories shoot all over the place. If time allotted, one reason that forecasted trajectories may deem themselves pertinent could be if a team qualifies for the playoffs, how will the score differential of the last game of the season, or the second to last game of the season effect the score differential of the beginning of the play-offs.

Conclusion

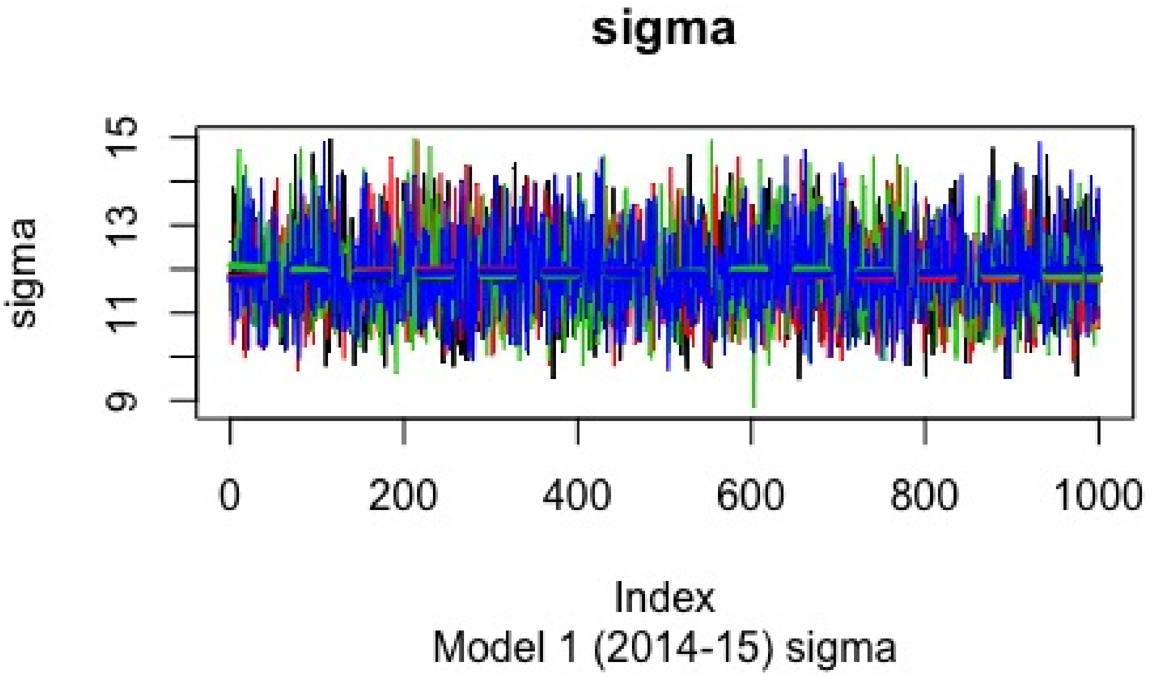
Fitting separate splines models for subsets of data allowed for a direct comparison between Doc Rivers and Brad Stevens. From our results, we can conclude that Doc Rivers was a more traditional coach, in the sense that the Boston Celtics, as one would expect, performed better against lower ranked teams. We can also conclude that Brad Stevens is not a traditional coach, because the Boston Celtics performs better against higher ranked teams, contrary to what one would expect. However, the three splines models gave us a quantitative way to assess who is superior: Doc Rivers, as well as a way to predict expected score differential for the Celtics given the ranking of the opposing team.

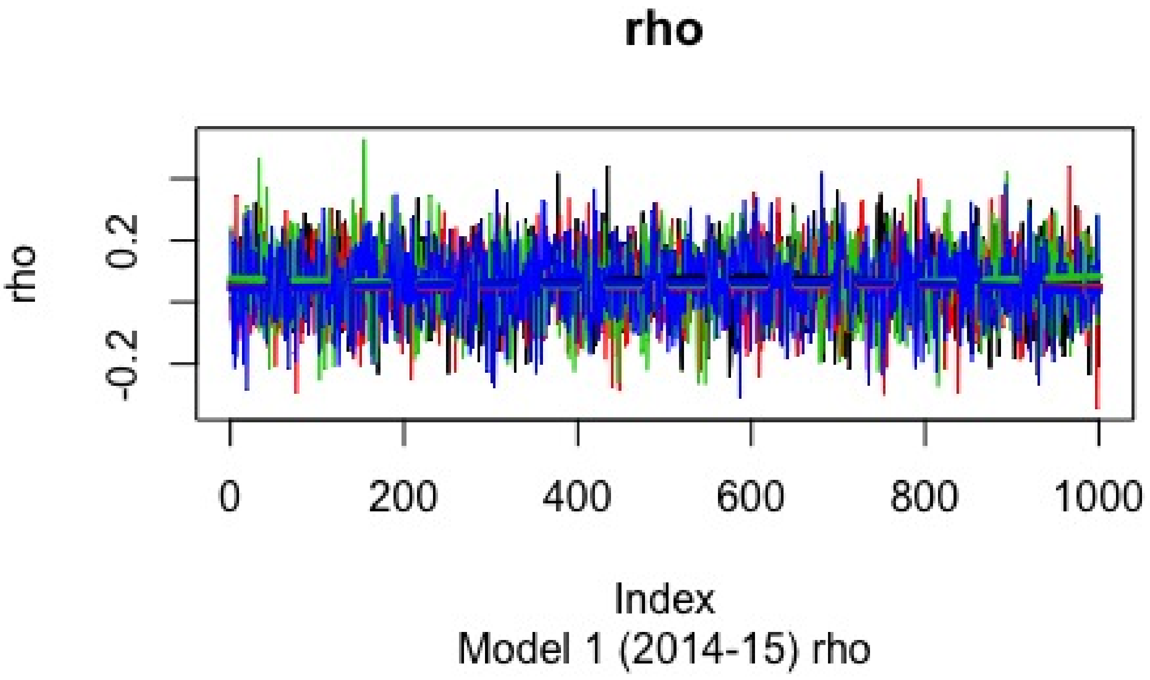
Fitting a Bayesian AR(1) process in order to estimate y_t, the score differential, proved to be a cumbersome task, but served its purposed in exhausting the model fitting process. We saw that Bayesian estimates of σ and ρ varied slightly from the frequestist estimates. Additionally, by fitting trajectories on a poorly fit AR(1) process did not display much information, except for the fact that the trajectories are quite unpredictable when your model does not fit any of the initial assumptions (such as the auto-correlation function plot). For future research, we would look into fitting much more sophisticated models, adding covaraites, as they were not added for the sake of time and for the limited scope of this subject in class.

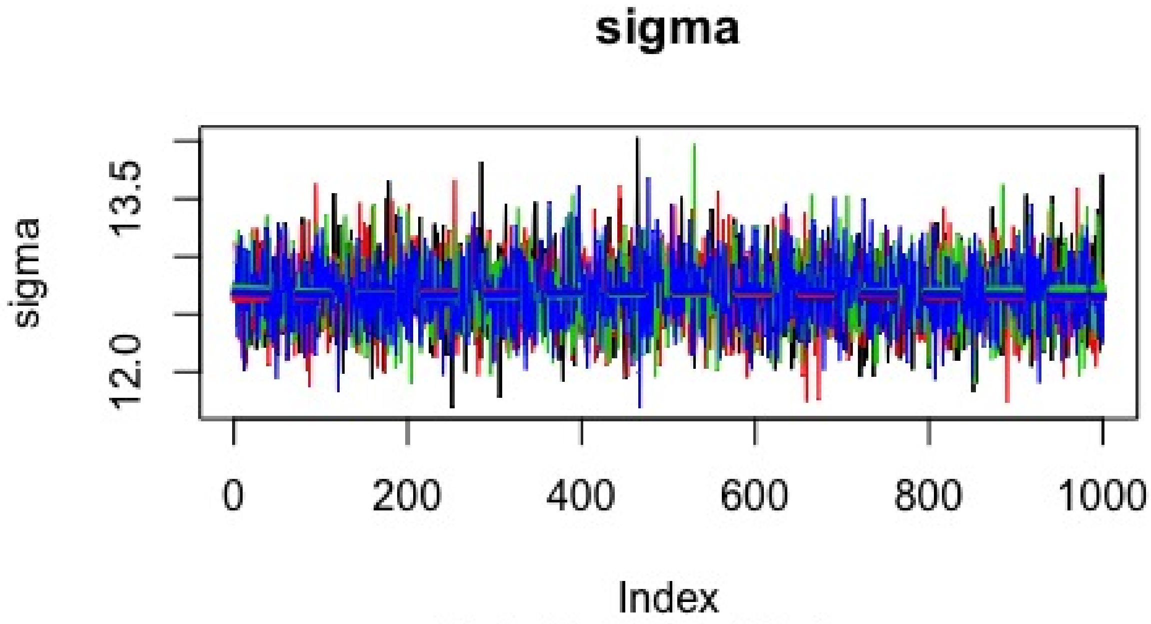
Appendix

Traceplots

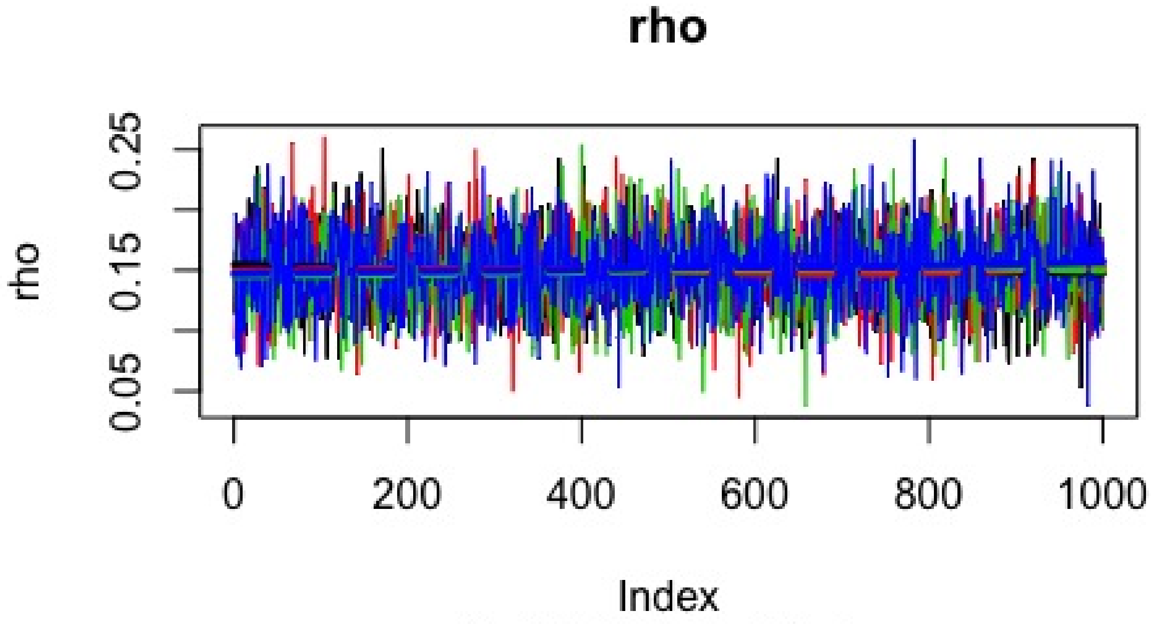
Traceplots for Time Series Models 1 (2014-2015 season) and Model 2 (2003 - 2015 season)







Model 2 (2003-15) sigma



Model 2 (2003-15) rho

JAGS Code

Splines JAGS code

```
model{
  for (i in 1:n){
    y.i[i]~dnorm(mu.i[i],tau.y)
    mu.i[i] <- inprod(B.ik[i,], alpha.k)
  }
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif(0,3)
  for (k in 1:K){
    alpha.k[k] ~ dnorm(0, 0.01)
  }
}
```

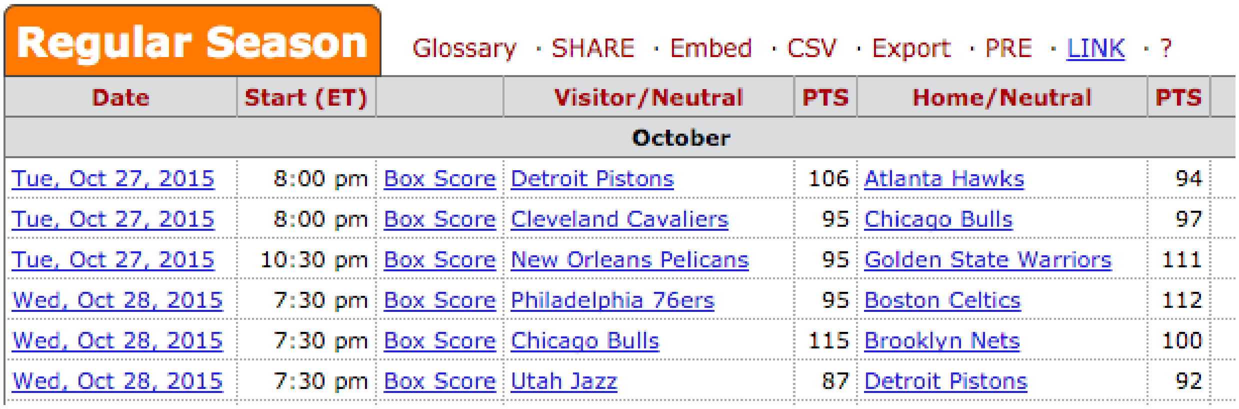
Time Series JAGS code for Model 1

```
model{
  y.t[1] ~ dnorm(0, tau.stat)
  for(t in 2:ngames){
    y.t[t] ~ dnorm(yhat.t[t], tau)
    yhat.t[t] <- rho*y.t[t - 1]
  }
  y.p[1] ~ dnorm(yhat.p[1], tau)
  yhat.p[1] <- rho*y.t[ngames]
  for(p in 2:P){
    y.p[p] ~ dnorm(yhat.p[p], tau)
    yhat.p[p] <- rho*y.p[p-1]
  }
  tau.stat <- (1 - pow(rho, 2))/pow(sigma, 2)
  tau <- pow(sigma, -2)
  sigma ~ dunif(0, 15)
  rho ~ dunif(-1, 1)
}
```

Time Series JAGS code for Model 2

```
model{
  y.t[1] ~ dnorm(0, tau.stat)
  for(t in 2:nceltcomp){
    y.t[t] ~ dnorm(yhat.t[t], tau)
    yhat.t[t] <- rho*y.t[t - 1]
  }
  y.p[1] ~ dnorm(yhat.p[1], tau)
  yhat.p[1] <- rho*y.t[nceltcomp]
  for(p in 2:P){
    y.p[p] ~ dnorm(yhat.p[p], tau)
    yhat.p[p] <- rho*y.p[p-1]
  }
  tau.stat <- (1 - pow(rho, 2))/pow(sigma, 2)
  tau <- pow(sigma, -2)
  sigma ~ dunif(0, 15)
  rho ~ dunif(-1, 1)
}
```

Screenshot of website



Date	Start (ET)		Visitor/Neutral	PTS	Home/Neutral	PTS
October						
Tue, Oct 27, 2015	8:00 pm	Box Score	Detroit Pistons	106	Atlanta Hawks	94
Tue, Oct 27, 2015	8:00 pm	Box Score	Cleveland Cavaliers	95	Chicago Bulls	97
Tue, Oct 27, 2015	10:30 pm	Box Score	New Orleans Pelicans	95	Golden State Warriors	111
Wed, Oct 28, 2015	7:30 pm	Box Score	Philadelphia 76ers	95	Boston Celtics	112
Wed, Oct 28, 2015	7:30 pm	Box Score	Chicago Bulls	115	Brooklyn Nets	100
Wed, Oct 28, 2015	7:30 pm	Box Score	Utah Jazz	87	Detroit Pistons	92