Determining Most Relevant Features in NBA Games and Comparing Linear Regression to Bayesian Regression

Introduction

We worked on this project because we were interested in determining which advanced statistics in the NBA are most important in determining a team's winning percentage. We also were interested in applying Markov Chain Monte Carlo (MCMC) simulations to a Bayesian regression model to evaluate its similarity to a general linear regression model.

Data and Variables

For the 2013-14 NBA season, we obtain data from NBA.com, basketball-reference.com, and nbaminer.com. We use the winning percentages for all 30 NBA teams as the outcome variable. We use the following as features: 3-Point Make Percentage (X3Ppct), 2-Point Make Percentage (X2Ppct), Assist Per Turnover Ratio (ASTpTO), Assist Ratio (ASTRatio), Steal Percentage (STLpct), Block Percentage (BLKpct), Turnover Ratio (TORatio), Personal Fouls Drawn Rate (PFDRate), Free Throw Attempt Rate (FTARate), Free Throws per Field Goal Attempt (FTpFGA), TurnoverPercentage (TOVpct), OREBpct, DREBpct, and the respective variables for each teams' opponents.

Code

```
library(R2OpenBUGS)
library ("arm")
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.10-1, built: 2018-4-12)
```

```
## Working directory is /home/jerry/Desktop/Bayes1_MarkovChainMonteCarlo_SDS384-master/NBAproject
```

```
modeldata<-read.csv("Teams2013_14.csv")

Winpct<-modeldata$Winpct
TORatio<-modeldata$TORatio
OppFTARate<-modeldata$OppFTARate
X2Ppct<-modeldata$X2Ppct
X3Ppct<-modeldata$X3Ppct
ASTpTO<-modeldata$ASTpTO
```

```
Opp3Ppct<-modeldata$Opp3Ppct
Opp2Ppct<-modeldata$Opp2Ppct

ORBpct <-modeldata$ORBpct
FtpFGA <-modeldata$FtpFGA
OppTOVpct <-modeldata$OppTOVpct

y<-Winpct
n <- length(Winpct)

previousseason <- list (n=n, y=y, X2Ppct=X2Ppct, TORatio=TORatio, ORBpct=ORBpct, FtpFGA=FtpFGA, OppTOVp

#Try JAGS below
library(rjags)
```

```
## Loading required package: coda
```

```
##
## Attaching package: 'coda'
```

```
## The following object is masked from 'package:arm':
##
##     traceplot
```

```
## Linked to JAGS 4.3.0
```

```
## Loaded modules: basemod,bugs
```

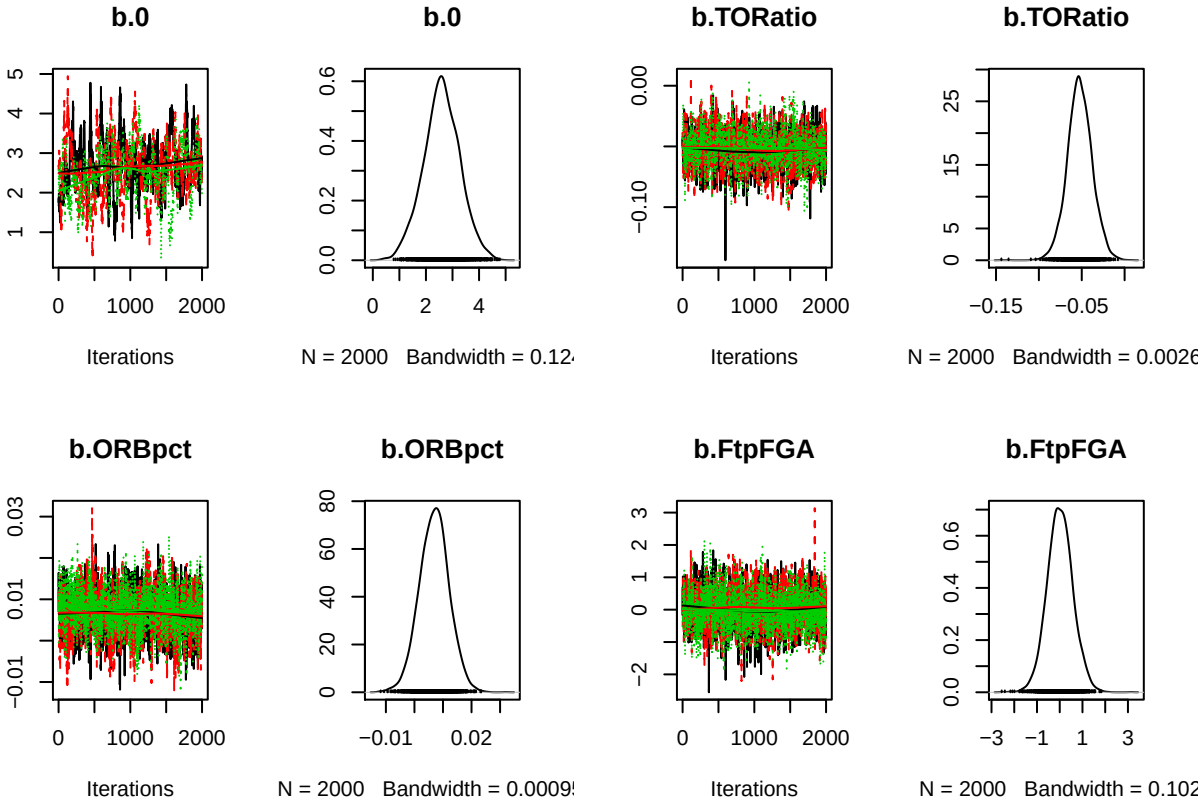
```
inits = function ()
{
  list('b.0' = rnorm(1, 0, 100), 'b.X2Ppct' = rnorm(1, 0, 100), 'b.TORatio' = rnorm(1, 0, 100),
       'b.ORBpct'= rnorm(1, 0, 100), 'b.FtpFGA'= rnorm(1, 0, 100), 'b.OppTOVpct'= rnorm(1, 0, 100),'b.O
}

model <- jags.model('modelNBAjags.txt', data=previousseason, inits=inits, n.chains=3)
```

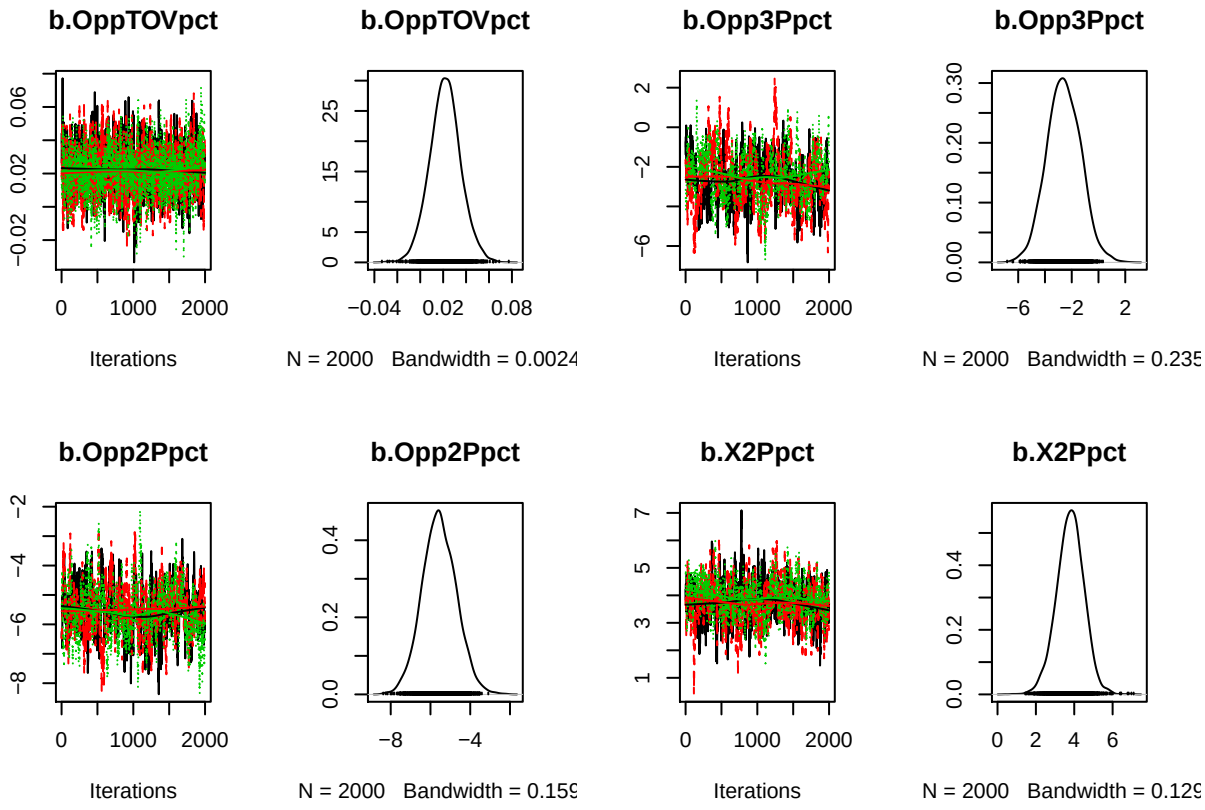
```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 30
##    Unobserved stochastic nodes: 9
##    Total graph size: 452
##
## Initializing model
```

```
update(model, 20000)
mcmc_samples <- coda.samples(model, variable.names=c('b.0','b.X2Ppct', 'b.TORatio', 'b.ORBpct', 'b.FtpF
post_samples <- jags.samples(model, variable.names=c('b.0','b.X2Ppct', 'b.TORatio', 'b.ORBpct', 'b.FtpF
post.b0<-as.mcmc.list(post_samples$b.0)
post.bTORatio<-as.mcmc.list(post_samples$b.TORatio)
```

```
post.bX2Ppct<-as.mcmc.list(post_samples$b.X2Ppct)
post.bORBpct<-as.mcmc.list(post_samples$b.ORBpct)
post.bFtpFGA<-as.mcmc.list(post_samples$b.FtpFGA)
post.bOppTOVpct<-as.mcmc.list(post_samples$b.OppTOVpct)
post.bOpp3Ppct<-as.mcmc.list(post_samples$b.Opp3Ppct)
post.bOpp2Ppct<-as.mcmc.list(post_samples$b.Opp2Ppct)
```

The trace plots below check that each feature variable converges, which verifies that our Bayesian regression model is valid.





Conclusion

Based on the data for the 2013-14 NBA season, we determined that the most relevant features were X2Ppct (Team's 2-point field goal percentage), TORatio (Turnover ratio), ORBpct (Offensive Rebound Percentage), FTpFGA (Free Throws per Field Goal Attempt), OppTOVpct (Opponent's Turnover Percentage), Opp3Ptpct (Opponent's 3-point Percentage), Opp2Ptpct (Opponent's 2-point Percentage), and ASTpTO (Assists per Turnover) were the most important features in determining a team's winning percentage

We also compared a general linear regression with a Bayesian regression model, and the models produced similar values for the feature coefficients except for the FTpFGA variable, which had a coefficient of .00763 for the linear regression model and .00377 for the Bayesian regression model. We expected the coefficients for the features to be similar for both models because we used non-informative priors for the Bayesian model, which should produce similar values as the frequentist methods