Predicting Winning in NBA Games Using Team Statistics

Jerry Kim and Wanyi Wang

SDS 384 Bayesian Statistical Methods

May 12, 2015

1. Introduction

During a National Basketball Association (NBA) season, large amounts of data are recorded during the game. There are "traditional" team and player statistics that are recorded in "box scores", such as the number of assists (AST), steals (STL), rebounds (REB), and field goal percentage (FG)%. However, with the recent rise of applying statistics and data science to the game of basketball, there are now new "advanced" statistics such as Effective Field Goal Percentage (EFG)%, Turnover Percentage (TOV)%, Rebounding Percentage (REB)%, and Free Throws Per Field Goal Attempts (FTpFTA), etc. Many people are interested in seeing how analytics can be used to predict players' performance and the results of games, including casual sports fans, those who engage in betting/gambling, statisticians, and NBA coaches and general managers. For example, Daryl Morey, the current general manager of the Houston Rockets, was the first NBA general manager to use analytics to make basketball decisions, such as deciding which players the Rockets should include in their team through trade, free agency, and the draft.

Previous studies have generated models to determine what basketball variables best predict the winning percentage of teams. For example, the statistician Dean Oliver came up with the "Four Factors" [1] in one of his studies. He pointed out that EFG%, TOV%, REB%, and FTpFGA were the four important factors that determined whether teams would win or lose games. However, these factors can be refined into more specific variables, which could give more details as to what best predicts whether a team will win or lose a game. For example, REB% could be further divided into Offensive Rebounding Percentage (ORB%) and Defensive Rebounding Percentage (DRB%). We wish to investigate our own linear regression model and compare our results with the “Four Factors”. In addition, we are interested in applying Markov Chain Monte Carlo (MCMC) simulations to a Bayesian regression model. Therefore, the first purpose of this project is to establish a more advanced model using general linear regression to investigate whether we obtain similar results as the Four Factors. The second purpose is to evaluate the similarity between frequentist and Bayesian estimates.

2. Data and Variables

The data for our project was collected from NBA.com, basketball-reference.com, and nbaminer.com. We use the winning percentages and statistics for all 30 NBA teams. The statistics we will use as predictors are: 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), Turnover

Percentage (TOVpct), OREBpct, DREBpct, and the respective variables for each teams' opponents. We use the winning percentage as the dependent variable

3. Models

3.1 Variable Selection from General Linear regression Model

We use the winning percentage and statistics for each team from the 2013-14 NBA season data to determine the multiple linear regression model. It is possible that all the predictors that we consider are strongly associated with teams' winning percentages, but it is more likely that the response is only related to a subset of the predictors. When trying to determine the relevant predictors, there are a total of 2^p models that contain subsets of p variables. This means that even for moderate p, trying out every possible subset of the predictors is infeasible. For example, if we try 26 different predictors, then with p=26, we have to consider 2^{26}=67,108,864 models. It is clearly not practical to try all of these possible models. Therefore, we use the Forward, Backward, and Stepwise Selection methods to determine which predictors are most relevant to predicting a team's winning percentage[2]. We briefly explain how these methods work: Forward Selection starts with no variables in the model, adds new variables to the model, and then tests these additional variables using a chosen model comparison criterion. This process is continued until no new predictors can be added. Backward Selection starts with all the predictors in the model and then removes the predictors with p-values greater than a specified criterion. Stepwise Selection is a combination of Backward Selection and Forward Selection. By combining the results generated from all three methods, we obtained the variables that are most relevant to predicting

a team's winning percentage. Two of the most common numerical measures, RSE and R^2, were used.

3.2 Bayesian Regression Model

The Bayesian regression model we use has the linear coefficients of the predictor variables as the parameters of interest, and the posterior distribution depends on what we use as the likelihood function.

4. Results

4.1 Variable Selection

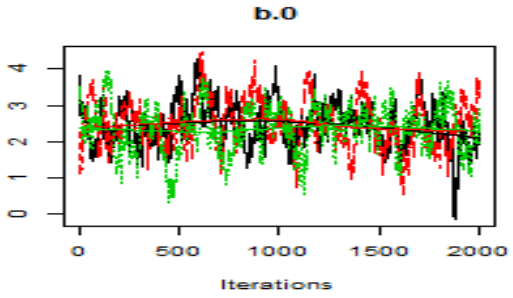
Combing the results from Forward, Backward, and Mixed Selection, we found that the most relevant predictors are X2Ppct, TORatio, ORBpct, FTpFGA, OppTOVpct, Opp3Ppct, and Opp2Ppct. All of these variables are contained in the "Four Factors" mentioned above, but our results have predictor variables that are more refined. For example, the shooting factor is measured using Effective Field Goal Percentage (EFG%), which is a combination of 2Ppct, 3Ppct, Opp2Ppct, and Opp3Ppct. From our results, only 2Ppct, Opp2Ppct, and Opp3Ppct are relevant. The rebounding factor is measured using Offensive and Defensive Rebound Percentage. Our results shows that only Offensive Rebound Percentage (ORBpct) has a significant impact on winning games.

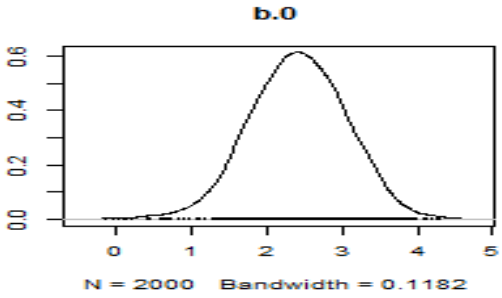
4.2 Coefficients from General Linear Regression

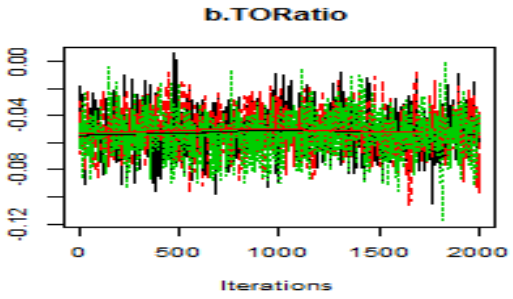
After determining the relevant variables, we determine that the values for the regression coefficients are 3.84, -.0528, .00722, .00763, .0219, -2.54, and -5.53 for X2Ppct, TORatio, ORBpct, FTpFGA, OppTOVpct, Opp3Ppct, and Opp2Ppct, respectively.

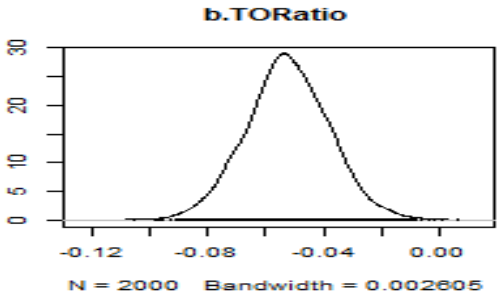
4.3 Bayesian Regression Using MCMC

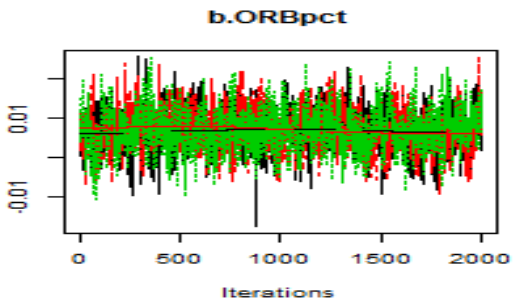
We use non-informative prior distributions for the parameters. That is, we use “flat” normal distributions. We then perform Bayesian regression using MCMC simulations using OpenBugs and rjags from within R. Our results show that the mean coefficient values for the predictors are 3.86, -.0523, .00755, .00377, .0227, -2.36, and -5.60 for X2Ppct, TORatio, ORBpct, FTpFGA, OppTOVpct, Opp3Ppct, and Opp2Ppct, respectively. We verify that our Bayesian regression model is valid by checking that the trace plots for each predictor variable converges. The plots are below in Figure 4.1. We also obtain values close to 1.00 for the Gelman diagnostics, with the highest values being 1.08 for Opp2Ppct and 1.06 for Opp3Ppct. Also, the largest in magnitude cross-correlation value we obtain is -.467 for the cross-correlation between Opp2Ppct and Opp3Ppct. However, our plots for the auto-correlation do not initially converge to 0 for Opp2Ppct, Opp3Ppct, and X2Ppct. However, after increasing the thinning, we see in Figure 4.2 the convergence in the auto-correlation plots that we expect

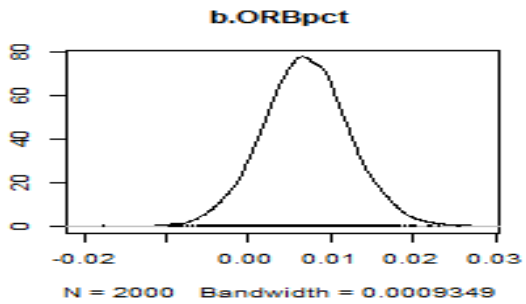

b.0
Iterations

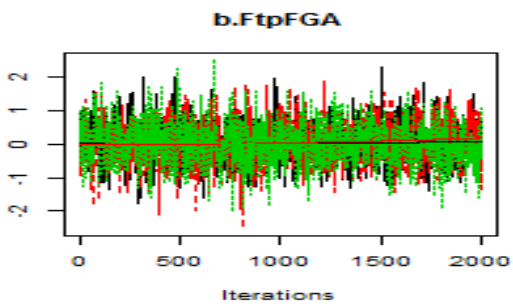

b.0
N = 2000 Bandwidth = 0.1182

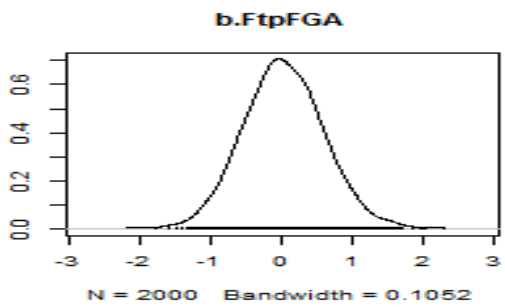

b.TORatio
Iterations


b.TORatio
N = 2000 Bandwidth = 0.002605


b.ORBpct
Iterations


b.ORBpct
N = 2000 Bandwidth = 0.0009349


b.FtpFGA
Iterations


b.FtpFGA
N = 2000 Bandwidth = 0.1052

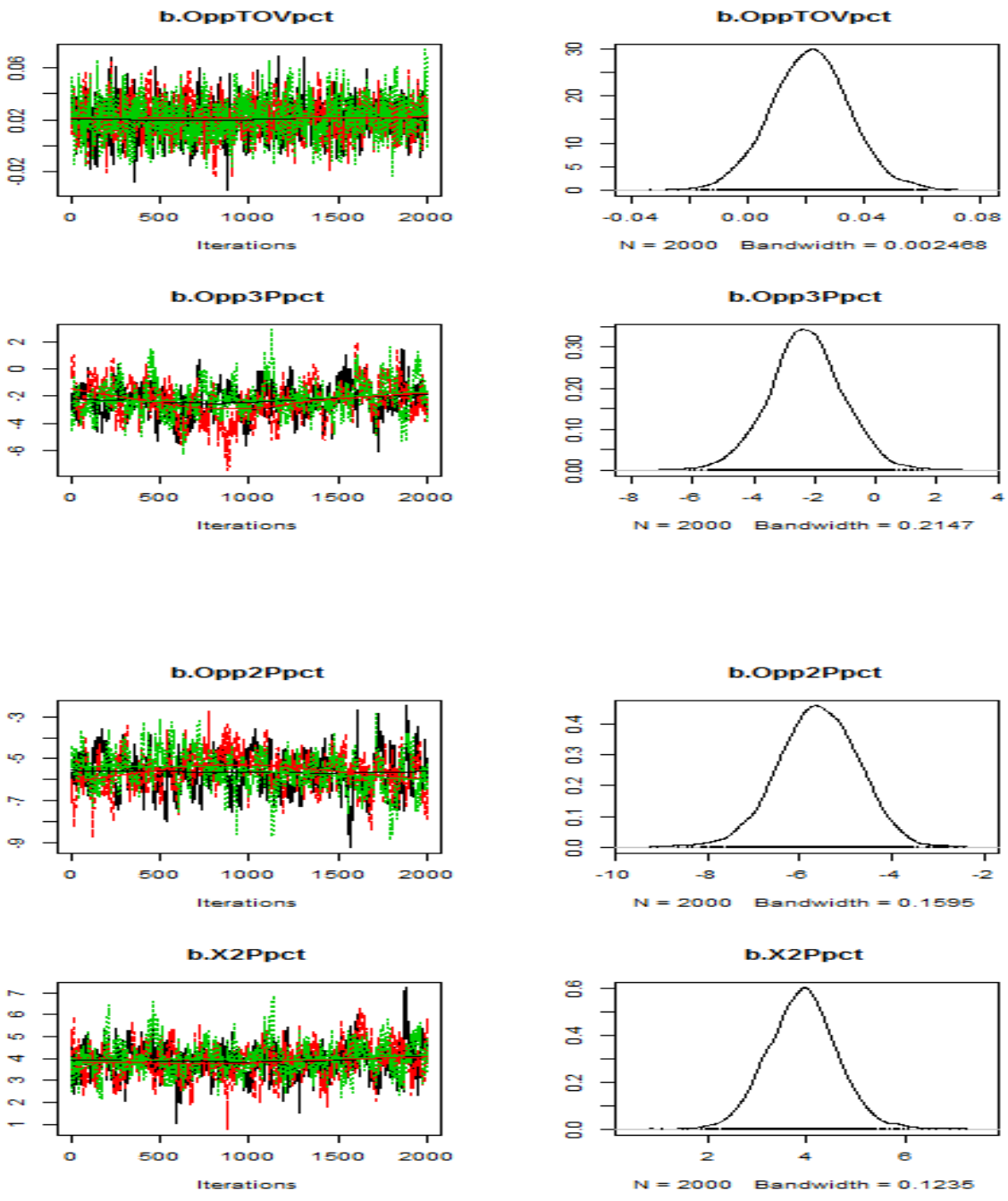


Figure 4.1: Trace and density plots for the predictor variables and the intercept. The green, black, and red traces indicate the 3 different chains. All the density plots resemble normal distributions

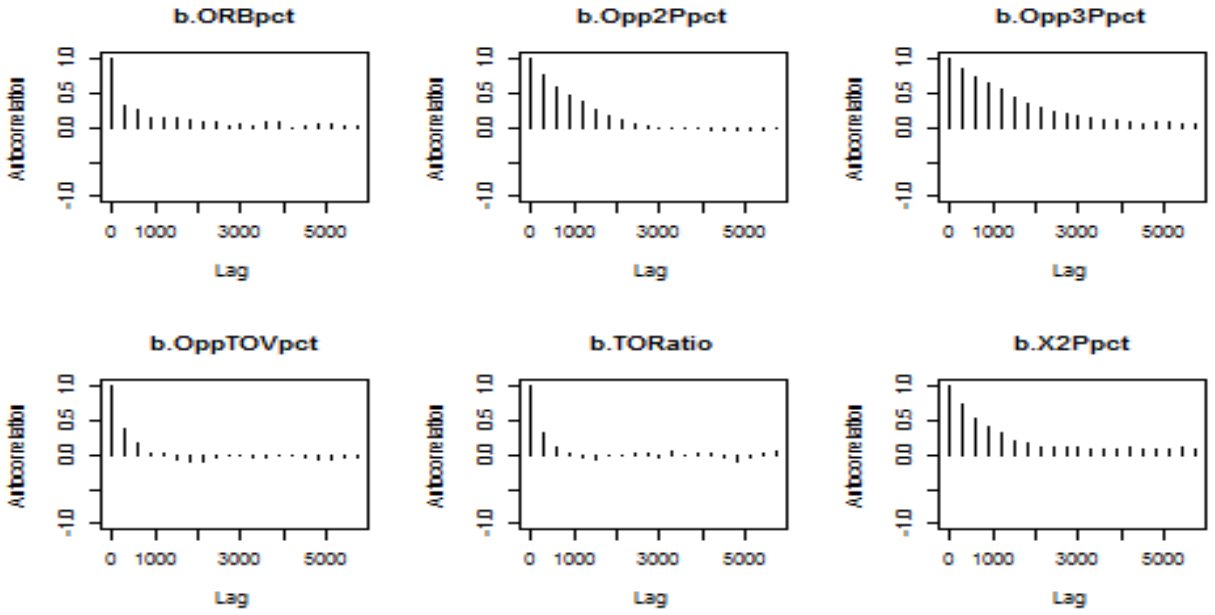


Figure 4.2: Auto-correlation plots for the predictor variables

5. Discussion

Based on the dataset in 2013-14 season, we found that the most relevant predictors are X2Ppct, TORatio, ORBpct,FTpFGA, OppTOVpct, Opp3Ppct, Opp2Ppct, ASTpTO. All of these variables are consistent with the "Four Factors" mentioned above, except that we include more precise variables in our model. Only 2Ppct, Opp2Ppct, and Opp3Ppct were included in the EFG%, and only ORBpct had a significant impact on the prediction of winning game instead of REBpct. In addition, we see that the general linear regression and Bayesian regression produced similar values for the coefficients of the predictors. However, the one exception is with the FtpFGA variable, which had a coefficient of .00763 for the linear regression model but value of .00377 for the Bayesian regression model. These similar values matches our expectations because we used

Bayesian methods with non-informative priors, which should produce similar values as the frequentist methods. For future research and to generate more accurate methods, there are several possibilities we could investigate. First, we could use informative priors for the Bayesian regression model. Second, we could utilize machine learning algorithms, such as the Random Forest algorithm to determine which predictors are most relevant. In addition, we could use Support Vector Machine with our predictor variables to try to predict the winners of future NBA games.

6. References:

1. Oliver, D. (2004). *Basketball on Paper: Rules and Tools for Performance Analysis.* Dulles, VA: Potomac Books
2. James, G., Witten, D., Hastie, D., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with application in R.* New York, NY: Springer