

5-7-2014

Estimation of NBA players' offense/defense ratings through shrinkage estimation

Kyongche Kang
Carnegie Mellon University

Follow this and additional works at: <http://repository.cmu.edu/hsshonors>

 Part of the [Statistics and Probability Commons](#)

This Thesis is brought to you for free and open access by the Dietrich College of Humanities and Social Sciences at Research Showcase @ CMU. It has been accepted for inclusion in Dietrich College Honors Theses by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

Estimation of NBA Players' Offense/Defense Ratings through Shrinkage Estimation

Kyongche Kang

May 7th, 2014

Acknowledgements

I would like to thank my advisor, Dr. Andrew C. Thomas, for introducing me to this amazing topic and the world of sports analytics. I appreciate your advice and patience that enabled me to push through and complete my thesis.

I dedicate my thesis to my family, and especially my parents who have worked and sacrificed so much to provide me an outstanding education. I will never forget what you have gone through. From the bottom of my heart, I thank you.

Abstract

The standard plus/minus model for rating NBA players combines the offensive and defensive capabilities of each player into a single metric. While this is convenient for the sake of summary, it makes it difficult to isolate the particular contributions that a player makes to either effort. Although adjusted plus/minus and other methods are proposed to address this, given a relatively large number of players against the number of events observed in one season, estimates are subjected to high variance. To correct for this, we construct a penalized regression model that identifies the specific offensive and defensive contributions of each player on each possession, and tune the model using L2-regularization method to optimize its predictive power. It overcomes the limitations of simple and adjusted plus/minus by incorporating offensive and defensive effects separately, and the shrinkage term controls for high variance of the estimates. Furthermore, our model captures net home court advantage on offense, and estimate players' contributions in offense and defense. Finally, we demonstrate application of our method by simulating unseen matches to correctly predict their outcomes.

Contents

- 1 Introduction** **5**

- 2 Previous Approaches** **7**
 - 2.1 Simple Plus/Minus 7
 - 2.2 Adjusted Plus/Minus 8
 - 2.3 Regularization Methods 11

- 3 Analysis** **13**
 - 3.1 NBA Season 13
 - 3.2 Data 14
 - 3.3 Model 17

- 4 Results** **22**
 - 4.1 Team Ratings 22
 - 4.2 Player Ratings 26
 - 4.3 Discussion 31
 - 4.4 Game Outcome Prediction 31

- 5 Conclusion** **34**
 - 5.1 Limitations 35
 - 5.2 Future Work 35

Chapter 1

Introduction

In many sports, we are curious about how much players contribute and influence the outcome of the game. It is very easy to estimate and measure players' contributions in individual sports, such as golf or tennis; while holding the other factors constant, there is only one person's performance against the opponent in the game that determines the outcome. It is also easier to compare, rank and calibrate players based on their abilities. However, in team sports like basketball, it is trickier - there is more than one player on the court per team during the game, and roles of players are not just fixed to offense or defense. Some players' roles are not just confined to offense, but they also participate in defensive or supporting roles. These types of roles may not have direct influence on the outcome in a measurable way. In this sense, basketball is a sport that it is harder to measure players' contributions than other sports; it is fast paced, where players are alternating in offensive or defensive actions within a split second. Although players have specific positions, their plays are not solely confined to offense or defense, but mixture of both, as well as indirect offensive or defensive supporting plays that are not easily quantified.

In this analysis, we attempt to isolate player effects from noise, and correctly define their contributions to determine how their performance on the court contribute to likelihood of scoring a basket, thus influencing the probability of winning. We introduce penalized logistic regression, particularly ridge regression using L2-regularization, which effectively describes each player's contribution during a season compared to an average player at that position.

In the following Chapter 2, we describe previous and popular approaches that are used to measure players' abilities in basketball. In Chapter 3, we discuss our data, and methods for cleaning and summarizing which enable us to apply them in our analysis. In Chapter 4, we introduce our methodologies to correctly measure team and player abilities, and player ratings results from our analysis. Finally in in Chapter 5, we present our conclusion, application, limitations and future work.

Chapter 2

Previous Approaches

2.1 Simple Plus/Minus

Plus/minus was first used in ice hockey; in 1950, Montreal Canadiens of National Hockey League (NHL) was the first team to track the plus/minus of its players. It was popularized by the NHL coach Emile Francis in 1960s, and the NHL started officially compiling the statistic for the 1967-1968 season. It is widely adopted as we know today, due to the lack of defensive measures. Plus/minus is a metric that measures team's overall performance, as well as players' offensive and defensive abilities. In team sports like ice hockey or basketball, scores and other statistics reflect offensive abilities, but often defensive abilities are not easily quantifiable. It is adopted and developed further in basketball. Roland Beech (2003) first applied this concept to basketball and it became a popular method used in basketball to measure player contribution on the court. The National Basketball Association (NBA) adopted it as part of official boxscore statistics. Plus/minus is a metric that looks at how teams perform with a certain player on the court, how they perform with a certain player off the court, and calculates the overall impact that player has on team success (Sporting Charts). Here is an example on how to calculate simple plus/minus:

LeBron James came onto the court when the Miami Heat was up against the Oklahoma City Thunder by 5 points. Here, James plays and his team scored 10 points, leading by 15 points. James clashed with Kevin Durant and incurs an injury, which forced him to be substituted. Then, James's plus/minus is +15, score difference after James was on the court, minus +5, the original lead before James got on the court, for +10 in total. On the other hand, Durant came on at the same time as James did, and he also incurred an injury after the clash. His team was down 5 points when he got on the court, and when he left, the Thunders was down by another 10 points. So Durant's plus/minus would be -15 minus -5, which is -10 in total. Overall, James receives +10, while Durant receives -10 plus/minus.

Despite its popularity, simple plus/minus statistics is flawed as a method of rating players. As it is entirely based on the score differences, the rating is not indicative of individual player's performance. It only tracks the differences in score while players are on court, and many of the statistics can be attributed by team's or the opponent team's abilities. It may be the case that a strong offensive player in a weak team may receive the same plus/minus as an average offensive player in a strong team, because other team members' contributions can drive the outcome of the game. It is hard to isolate each player's contribution from this simple statistic.

2.2 Adjusted Plus/Minus

In order to account for the flaws in simple plus/minus, Rosenbaum (2004) proposed linear regression-based adjusted plus/minus approach. Rosenbaum attempted to address the problem of substitution of players. Due to the nature of basketball, players are substituted frequently. This makes it harder to pin down individual's contribution, because substitutions change the dynamic of the game, and each player's contribution changes when one of the team members changes. In order to account for this, Rosenbaum grouped observations in

intervals marked by substitution. Each observation is an independent event in a game, with some scoring opportunities and also a period in time when no substitution is made.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon \tag{2.1}$$

Y corresponds to margin, which is defined as a difference between home team’s score points per possession and away team’s score points per possession, multiplied by 100. This is equivalent to points difference per 100 possessions. Each independent variable from $X_1 \dots X_p$ represents a NBA player, +1 for playing home game, and -1 for away game, and 0 not playing. The outcome of this model is the difference of points scored by home and away teams, divided by the time interval elapsed. By fitting a linear regression model, the intercept represents the average home court advantage across all teams, and the beta coefficients represent the contributions of players compared to that of a reference player to the outcome of the game.

Table 2.1: Data setup for Adjusted Plus/Minus (Witus, 2008). Margin as the difference between home and away teams’ points per 100 possessions, Unique NBA Players represented as Player 1 through Player P for P number of NBA players - Home Player on court as +1, Away Player on court as -1, Not on the court 0.

Margin	Player 1	Player 2	...	Player P-1	Player P
65	1	1	...	-1	0
200	1	1	...	0	-1
-200	1	1	...	-1	0
⋮	⋮	⋮	⋮	⋮	⋮
-100	-1	-1	...	0	1
300	-1	-1	...	0	1
47	-1	-1	...	0	1
-66	-1	-1	...	1	0

After Rosenbaum's initial proposal, adjusted plus/minus was further developed by Lewin (2007), Witus (2008), and Ilardi and Barzilai (2008). Lewin and Witus followed Rosenbaum's approach to publish the adjusted plus/minus for the following 2005-2006, 2006-2007 seasons by Lewin, as well as 2007-2008 seasons by Witus. In particular, Witus describes steps in details on how to follow Rosenbaum's adjusted plus/minus. He provides instructions for readers on how to prepare the data, and model fitting in R in his blog CountTheBasket.com.

In the season 2008-2009, Ilardi and Barzilai (2008) modified Rosenbaum's approach by incorporating multiple seasons, and they published adjusted plus/minus from 2007-2008 to 2011-2012 season. Rosenbaum's adjusted plus/minus aggregates player abilities in a single metric - in Ilardi and Barzilai's modification, they separate offensive and defensive abilities. Instead of fixing the margin as the difference between home and away team points per possession, they redefine the margin as the difference between points of offense team and those of defense team per 100 possessions; it alternate from difference between home and away, or between away and home teams based on which team is on offense. The player indicators also changes from fixing +1 to home and -1 to away team players, to allowing indicators to vary by offense/defense possession: offense variable ranges from $\{+1;0\}$, and defense variable from $\{-1;0\}$, when they are on the court. They also include home offense indicator: whether home team is on offense or not. This indicator captures net home court advantage on offense, and how that contributes to player performance. Ilardi and Barzilai's new adjusted plus/minus incorporates desirable features that are not captured in Rosenbaum's approach. It isolates individual players' abilities by offense and defense, and also takes into account of home team's advantage.

2.3 Regularization Methods

Adjusted plus/minus overcomes bias and skewed measures of simple plus/minus; but this regression method is prone to high variance and multicollinearity, due to relatively large number of predictor variables against the number of observations in a season. There are about 500 different NBA players per season, but typically the number of observations per each player in a season is under 1,000 (Ilardi and Barzilai, 2008). This is likely not sufficient to give accurate estimates. It also has multicollinearity issue, as in most of basketball games, teams tend to put the same group of players on the court at the same time. Linear regression approach fails to properly address players with very small number of samples - their few plays have high variance in their plus/minus scores.

Alternative to adjusted plus/minus, Thomas et al. (2012), and Gramacy et al. (2013) introduce methods for modeling partial effects of each player in the game, which control for abilities of teammates and opponents, and other desired factors. Specifically, using regularization and shrinkage can yield more robust estimates of players' contributions in games with sparse individual player observations, like in ice hockey. As a result, estimates are indicative of individual players' true abilities compared to reference player directly related to the outcome of the game, rather than the score differences that players are correlated with during the game.

Thomas et al. (2012) model the scoring rate for each of home and away team as its own semi-Markov process, using a Cox proportional hazard function. This method is able to calculate offensive and defensive ability of each player simultaneously, instead of modeling a single outcome for offense/defense.

Gramacy et al. (2013) introduce a regularized logistic regression for evaluating player contribution, and applying these estimates as a prior distribution to simulate Markov Chain Monte Carlo, and to estimate a posteriori distribution (MAP), using Bayesian approach.

Thus, they are able to identify star players, whose contributions stand out in the process.

Plus/minus itself fails to give clear measures of true player abilities. It is a metric that measures a player's marginal effect in the score during the game. Many factors are associated with scoring, and it is not just a player's ability that drives the statistic. A player's teammate abilities, and his opponent abilities are very important in scoring a basketball, but they are not constant across players, nor adjusted plus/minus controls for these effects. It is hard to isolate individual player contribution, and also to compare across players.

Our new approach is motivated by the previous approaches mentioned above. Gramacy et al. (2013) demonstrate effectiveness of logistic regression in modeling players' abilities, compared with linear regression by Rosenbaum (2004). This is especially effective in basketball, where frequent scoring events take place. Ilardi and Barzilai's (2008) modification of adjusted plus/minus incorporates offense/defense ratings in adjusted plus/minus, rather than overall ratings; they also incorporate net home court advantage on offense in the model. Thomas et al. (2012) introduced shrinkage in the model, to address the issue of outlier players who might have a few, but exceptional plays. Our method incorporates core features of each method, to produce a model that accurately estimate the NBA players' offensive and defensive abilities through scoring a basket. Shrinkage is particularly important for this method, as we incorporate a relatively large number of predictor variables.

Chapter 3

Analysis

3.1 NBA Season

In a season, there is a total of 82 games in a regular season, and 16 games in the season play-offs per team. There are 30 teams across North America, and they play against every other team during the season. Teams are organized into two conferences, Eastern and Western, and each conference is also consisted of three divisions. These are organized by geographical location. Although teams play against each other, only the top 8 teams in each conference proceed to play-offs. Play-offs are elimination tournament among 16 teams, and the top team from each conference plays the finals to determine the winner of the season. We are interested in the NBA season 2011-2012. This season, however, was shortened to 66 games per team due to lockout of the collective bargaining agreement between the team owners and the players. In our analysis, we take into account of 66 regular season games for our results and exclude the play-offs - not every team plays in the play-offs, and this results in different number of observations per each team. This allows our results to be consistent and comparable across teams. To summarize the season, Miami Heat and Oklahoma City Thunder played in the finals, and Miami Heat won the championship. LeBron James won

the Most Valuable Player (MVP) award, and Tyson Chandler from New York Knicks won the Defensive Player of the Year in 2012.

3.2 Data

The NBA 2011-2012 data are obtained from publicly available source, BasketballValue.com, which is a website maintained by Barzilai. There are various statistics and information about the overall season, matches, teams and players, but we are particularly interested in Play-by-Play and Match-ups datasets. Play-By-Play data provides descriptions of events that happened during the game, from the beginning to an end of each game played in a season. The descriptions provide details on teams and players involved in the action, types of events, and the updated scores when the ball is scored in the basket. Match-ups table provides various statistics, such as simple plus/minus, number of possessions, offensive/defensive rebound rates by home and away teams, as well as time intervals. It also provides the names of home/away players on the court during that time intervals. Both datasets contain unique game ID, identifying observations with individual matches, and time remaining in minutes when the event was observed.

In order to incorporate these information together, we combine the two data tables of Play-By-Play and Match-ups. Both sets have time remaining in minutes when the events occurred. We join the two datasets by two keys, one with time remaining and unique game ID. Match-ups dataset contains a fewer number of rows, and the rows are duplicated if they are within the time frame of the events in Play-By-Play. The event entry in Play-By-Play only contains plain text descriptions on what happened during the game. For example,

“[MIA 4-6] James Layup Shot: Made (2 PTS) Assist: Wade (1 AST)”

“[MIA] James Rebound (Off:0 Def:4)”

“[OKC] Durant Turnover : Bad Pass (3 TO) Steal:James (1 ST)”

The information in the brackets contain the team in action, updated scores in case of scoring event, and description of what happened with the names of players involved. The name of the player in action appears in the beginning of the sentence after the brackets of team information, and additional information on different player's involvement also appears at the end, such as steal, assist or block, which affects the outcome of the action. Following verbs comprise types of actions described by the event entries:

Offense	Defense
Shot, Free Throw, Rebound (Offensive)	Foul, Rebound (Defensive), Turnover

The one sentence description contains all the information that we need. Using regular expression, we extract the following information and coerced them into our combined dataset:

$$\begin{aligned}
 ScoringEvent &= \begin{cases} +1, & \text{scoring event} \\ 0, & \text{otherwise} \end{cases} \\
 HomeOffense &= \begin{cases} +1, & \text{home team on offense} \\ 0, & \text{otherwise} \end{cases} \\
 Offense &= \begin{cases} +1, & \text{players on the court while their team on offense} \\ 0, & \text{otherwise} \end{cases} \\
 Defense &= \begin{cases} +1, & \text{players on the court while their teams on defense} \\ 0, & \text{otherwise} \end{cases} \quad (3.1)
 \end{aligned}$$

Home Offense indicates whether home team was on offense (net home court advantage on offense). +1 is assigned to events when home team is on offense. Scoring Event indicates whether a ball is scored in the basket. +1 is assigned to scoring event. Player Offense and

Defense are matrices that indicate whether players are on offense or defense. At each event observation, one team is on offense, and other team is on defense. Individual players are represented as columns. For Player Offense, players on the court whose team is on offense are assigned with +1, and 0 if they are not on the court, or not part of the team. Similarly, players whose team is involved in defense are assigned +1 and 0 otherwise for Player Defense.

The player dataset contains 518 NBA players during 2011-2012 season. As observed, the event entries and many other datasets where players are mentioned, only their last names are used. Hence, this creates confusion in exactly which players are involved, as some players have the same last names. For example, there are two players with the last name Gasols in NBA, Marc and Pau Gasol, in Memphis Grizzlies and Los Angeles Lakers respectively. There are also three players with the last name James, LeBron James in Miami Heats, Mike James in Chicago Bulls, and Damion James in New Jersey Nets. In order to recognize and separate players, team code is appended to the player's name. Among 518 players, there are duplicate records of the same player. Those players got traded to a different team in the middle of the season. As we take into team effect into account, we treat it as unique players. A player's performance is not independent of his teammates, and we want to control for that effect.

For this analysis, we consider events that are directly related to the outcome of the game. There are some event entries that are unclear or not specifically related to offensive or defensive actions. In particular, violation is one that is less severe than foul, but there are more violation rules than foul rules, and there is not enough information given to indicate whether they are committed by players on offense or defense. Such events like 'Violation', 'End Game' or 'Time Out' are excluded from our data. On the other hand, foul is an important indicator in describing defensive plays, as most fouls are committed by players on defense.

Furthermore, we compress observations in intervals to reflect the true nature of the bas-

ketball game. Rosenbaum’s interpretation of grouping observations to intervals is marked by when a player substitution is made. Our interpretation is slightly different - we recognize that basketball is a game with quick exchanges between offense and defense. There are various events that are leading up to one outcome of the game under one team’s possession. That outcome, whether successful or not, ends when the ball possession changes from one team to the other. Hence, we define our intervals from the point when a team gains possession, until it finishes its action, and then possession changes to the other team. Given frequent and large numbers of scoring events and observations available, we are not losing in important scoring information through this process.

3.3 Model

We model the probability of scoring, given the home team effect and player presence, to estimate players’ contributions in offense and defense. Given $p + 1$ number of variables (with 1 home team effect and p NBA players in a season, for each observation i), we formulate a logistic regression as the following:

$$\log \left(\frac{p(x_i)}{1 - p(x_i)} \right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p+1} x_{ip+1} \quad (3.2)$$

We simplify the expression so that $x_i = x_{i1} + \dots + x_{ip+1}$. Solving for $p(x_i)$, we obtain the following:

$$p(x_i) = \frac{e^{\beta_0 + x_i \beta}}{1 + e^{\beta_0 + x_i \beta}} \quad (3.3)$$

In our model, the response variable $p(x_i)$ is scoring event probability, and the predictor variables are net home court advantage on offense (x_{i1}), player offense and player defense ($x_{i2 \dots ip+1}$). The intercept β_0 is the baseline effect. Our dataset indicates that about 55% of events are scoring event. Thus, we expect the baseline effect to be around 0, about 50%

chance of scoring, controlling for net home court advantage and player effects - we agree that just by a chance, a shot is made 50% of the time. The model enable us to model the players' contributions to offense and defense, as well as net home court advantage by their partial effects on log-odds of scoring in the basket. Our data are arranged as the following:

$$Y_{Scoring} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \quad X_{home} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}$$

$$X_{OFF} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \dots & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & \dots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad X_{DEF} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \dots & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 & 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & 1 & \dots & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.4}$$

Each row represents an observed event. In player offense and defense, 10 players are on the court at each event. Only a group of 5 players are in each offense and defense matrix at each given time. In the model, predictor X is a matrix that concatenates $X_{home}, X_{OFF}, X_{DEF}$.

In order to address the issues from linear regression by Rosenbaum (2004), we propose penalized regression to be applied to our data. Given a relatively large number of predictor variables (number of players) against the sample size (number of observations), our method

attempts to adequately address issues of high variance and multicollinearity as mentioned earlier. In regression, we use Maximum Likelihood Estimation to estimate the unknown parameter, β . The formula is the following:

$$\text{argmax } l(\beta|X) \quad (3.5)$$

Specifically, we apply logistic regression, where response variable has binomial distribution. We define the log-likelihood function $l(\beta|X)$ as the following:

$$L(\beta|X) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \quad (3.6)$$

$$l(\beta|X) = \log(L(\beta|X)) = \sum_{i=1}^n y_i \log p(x_i) + (1 - y_i) \log 1 - p(x_i) \quad (3.7)$$

Combining the result from 3.3, we maximize the log-likelihood function to find the estimates of β by taking the derivative of the following equation with respect to β :

$$\sum_{i=1}^n -\log(1 + e^{\beta_0 + x_i \beta}) + \sum_{i=1}^n y_i (\beta_0 + x_i \beta) \quad (3.8)$$

For each β_k , we solve for the following:

$$\sum_{i=1}^n (y_i - p(x_i; \beta_0, \beta)) x_{i,k} \quad (3.9)$$

However, we are interested in applying the shrinkage term to address the issues in the normal method. A penalized regression involves maximizing the following function:

$$\text{argmax } l(\beta|X) - \lambda P(\beta) \quad (3.10)$$

Where $P(\theta)$ is a penalty term. Maximizing a likelihood function is equivalent to minimizing

the loss function. We formulate the following for a regression with penalty term:

$$\operatorname{argmin} \sum_{i=1}^n (y_i - \beta_i x_i)^2 + \lambda \sum_{i=1}^n P(\beta_i) \quad (3.11)$$

The penalty term varies depending on the type of regression. Three types of penalty terms are represented as the following:

Type	Penalty Term
Lasso/L1	$\lambda \ \beta\ _1$
Ridge/L2	$\lambda \ \beta\ _2^2$
Elastic Net/L1+L2	$\lambda_1 \ \beta\ _1 + \lambda_2 \ \beta\ _2^2$

(3.12)

To write the penalty terms more elaborately,

$$L1 = \lambda \sum_{i=1}^n |\beta_i|$$

$$L2 = \lambda \sum_{i=1}^n \beta_i^2 \quad (3.13)$$

The main difference between L1 (Lasso) and L2 (Ridge) regularization is that due to the nature of penalty, the former causes some coefficients to be shrunk to zero exactly, while the latter doesn't. L1+L2 (Elastic Net) behaves similarly to L1, as it is a combination of L1 and L2. Since we are looking to assess every player's contribution, we want to keep the estimates of all the players. We impose L2 regularization for our penalized regression model. This requires selecting a specific value for shrinkage, λ . Optimal value of λ is chosen with automatic selection process of 10-fold cross-validation within R package, glmnet (Friedman et al. 2010). Logistic regression with ridge shrinkage value is fitted, and the optimal value

of λ is selected by minimizing deviance criteria of the model.

In sports analytics, offensive and defensive ratings are expressed as ω and δ . Following this convention, we make distinction on offensive and defensive ratings. β represents all the coefficients in the model, expressed in a matrix form. Our new model with L2-regularization is as follows:

$$\text{logit}(Y_{Scoring}) = \mu + X_{home}\alpha + X_{OFF}\omega + X_{DEF}\delta + \lambda\|\beta\|_2^2 \quad (3.14)$$

One thing to note is that Rosenbaum (2004) exclude players that did not play more than 250 minutes in the past two seasons, which became ‘reference players’ in his analysis. This type of players are typically ‘replacement’ players, substituting in for their teammates who are tired, or for the last few minutes of the game. We feel that those players are below-average, and having below-average players as reference does not give clear effects of players; Ilardi and Barzilai (2008) use 388 minutes as a cut-off. In our initial application, we notice that the shrinkage value is over-shrinking estimates of star players such as LeBron James or Kevin Durant. below-average players, who have substantially lower number of game appearances in the season, show greater contribution. Our approach to address this issue is that we group the players into a single player with a more rigorous cut-off: players with less than total possessions of 1000 are labeled as ‘replacement’. 1000 possessions translate to about 10 games, which are total of 480 minutes.

Chapter 4

Results

4.1 Team Ratings

We apply our method to estimate the teams' offense and defense ratings. We aggregate the constructed data to team level, and model with ridge regression. We find the optimal shrinkage through 10-fold cross-validation technique by minimizing deviance. The shrinkage value chosen is $\lambda = 0.072$. The result of cross-validation is shown in Appendix B. The choice of reference team in the model is important, as the interpretation of rating estimate is in relation to the reference team. We interpret offensive rating as a change in log-odds/probability of scoring when the team is on offense compared with the reference team. Similarly, defensive rating is interpreted as a change in log-odds/probability of scoring against when the team is on defense compared with the reference player - for example, the offensive rating estimate of 0.1 (52.5%) for team A refers to increase in log-odds(probability) of scoring for team A compared with the reference team while holding the effects of other teams constant. Hence, it is important to select an 'average' team, whose offensive and defensive abilities serve as the basis for comparing other teams' offensive and defensive abilities.

The reference team in our model is Phoenix Suns (PHX) as they are considered the most

‘average’ team with the equal number of wins and losses (33-33) in the season. Based on this interpretation, we recognize that positive, higher offensive rating estimates indicate strong offensive abilities as teams or players have higher chance of scoring on offense; similarly, negative, lower defensive rating estimates indicate stronger defensive abilities as teams or players decrease the chance of scoring against them on defense. We expect strong teams or players to have positive offensive ratings, and negative defensive ratings. The following results are the estimated ratings of the top 5 offense/defense ratings of the teams:

Top Offense Team	Top Defense Team
Denver Nuggets 0.086 (0.521)	Chicago Bulls -0.076 (0.481)
Utah Jazz 0.075 (0.519)	Boston Celtics -0.071 (0.482)
San Antonio Spurs 0.066 (0.516)	Philadelphia 76ers -0.066 (0.484)
Miami Heat 0.060 (0.515)	Dallas Mavericks -0.065 (0.484)
Oklahoma City Thunder 0.049 (0.512)	Miami Heat -0.064 (0.484)

(4.1)

According to team model, the baseline effect is -0.00057, and net home court advantage is 0.047. The baseline effect is close 0, which means a team has 50% chance of scoring - this agrees with the data. On average, home team has 51% chance of scoring, 1% advantage than baseline effect (0.047 log-odds) than away team at its home court while controlling for team effects.

With our data, we can also calculate the observed probabilities of teams scoring when on offense and on defense. These probabilities are calculated as the percentage of total number of scoring events observed against total number of offense or defense possessions for each team. This allows us to compare our estimates to observed data. The plot represents the relationship between observed probabilities against estimated probabilities by the regression in log-odds scale.

Team Rating Estimates: Offense vs. Defense

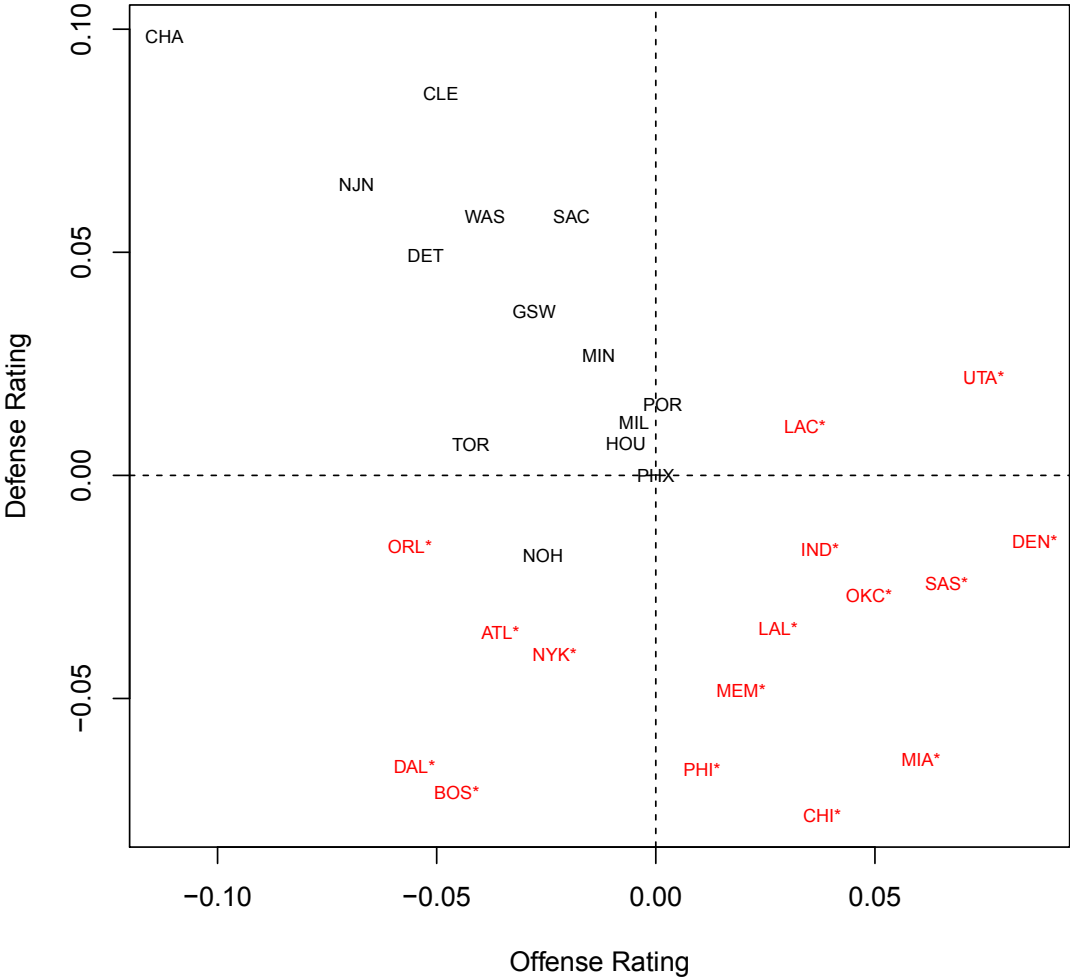


Figure 4.1: Estimated Teams' Offensive/Defensive Ratings. '*' indicates teams that made to playoffs in the season.

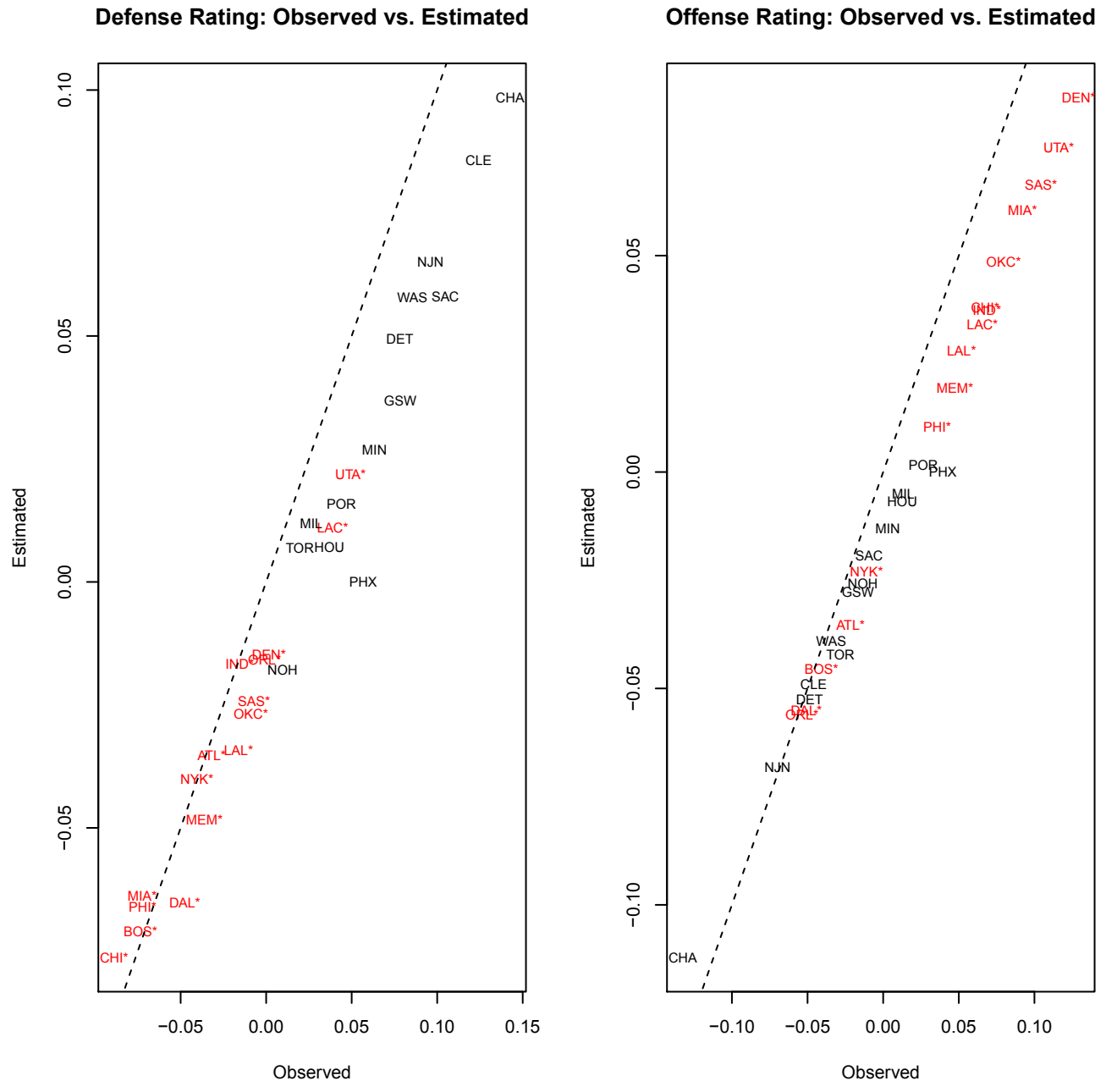


Figure 4.2: Observed Probabilities in Log-Odds Against Estimated Team Ratings

As teams are lining up in a straight line, we infer that the team ratings are indicative of team performance in the actual games. This means that our estimates seem to match the observed data. Towards the top of the plots, we see the shrinkage effects where teams are deviating from the straight line. We confirm that our method of estimating offensive/defensive ratings works on team level data.

4.2 Player Ratings

Applying the same method, we estimate partial effects of player contribution in the games through ridge regression. With 10-fold cross-validation, we obtain the shrinkage value $\lambda = 0.21$. The result of cross-validation is shown in Appendix C. The intercept is -0.0029 (50%), and the net home court advantage on offense is 0.0315. On average, home team player has 50.8% chance of scoring than away team player at its home court, controlling for player effects. We notice that we get different results for λ and estimates for intercept and net home court advantage on offense from the team model. The discrepancy comes from the different number of predictor variables in the models - player model has 520 variables, whereas team model has 31 variables. The shrinkage effect and net home court advantage on offense are adjusted to reflect this difference. The reference player in this model is Brendan Haywood from Dallas Mavericks - he showed the most average performance in offense and defense as center in the season (Blewis, 2012). The following are the top 5 offense/defense ratings of players:

Top Offense Player		Top Defense Player		
Manu Ginobili SAS	0.107 (0.527)	Thabo Sefolosha OKC	-0.086 (0.479)	
Ramon Sessions LAL	0.079 (0.520)	Jared Jeffries NYK	-0.081 (0.479)	(4.2)
James Harden OKC	0.072 (0.518)	Ben Wallace DET	-0.076 (0.481)	
Andre Miller DEN	0.062 (0.515)	Lavoy Allen PHI	-0.075 (0.481)	
LeBron James MIA	0.056 (0.514)	Ekpe Udoh GSW	-0.074 (0.482)	

The results for offensive ratings agree with what we expect, showing LeBron James as one of the top offensive players in NBA. James has 51.4% chance of scoring on offense. San Antonio Spurs had an extraordinary season, with the most number of wins (51 out of 66) games, and the least number of losses (15 out of 66) out of all the NBA teams. This explains why Ginobili, one of the top players for the SAS, topped all the other players in NBA with 52.7% chance of scoring on offense. Kevin Durant is another star player on the Oklahoma City Thunder (OKC), who led the team to the championship. He received an offensive rating of 0.028 (50.7%).

We note that Chandler, who won the Defensive Player of the Year award in 2012, is not among the top 5 defensive players. Instead, his teammate, Jared Jeffries, ranked the second. His presence on defense decreases the scoring chance by 47.9%. They are among the top defensive players in the New York Knick, but some critics argue that Chandler is a better defensive player than Jeffries (Zwerling, 2012). We suspect that the two players' observations are correlated, as they appear on the court often together. In our model, Chandler received a defensive rating of -0.0134 (49.6%). Although the top 5 defensive ratings of players may not identify some of the star defensive players, some ratings reveal the underrated players' true defensive abilities; Sefolosha's defensive contribution to the Oklahoma City Thunder also enabled his team to reach the championship, whose worth are not reflected on the

defensive stat-sheet contributions (Rinder, 2012). His presence on defense decreases the scoring chance by 47.8% against his team. We believe that ranking players by their ratings does not reflect their true abilities, as the rating's magnitude is in two decimal places, and in terms of probabilities, the difference between the best and worst ratings are not too far-off (standard deviation = 0.03). What is important is that the ratings have the appropriate signs: we expect positive values for offense, and negative values for defense, as their contributions increase or decrease the chance of scoring respectively. The majority of above-average players have positive offense and negative defense ratings; the below-average players, such as one labeled as 'replacement', have the opposite signs for offense and defense ratings. This means their presence has negative effects on the outcome of the game.

Given the number of players, it is difficult to see all the players in a single plot. We plot offensive and defensive ratings of players, whose total possessions are over 3,000. We focus in the players with exceptional abilities with positive offensive ratings and negative defensive ratings. This shows the top 15% of the NBA players. Similar plots for all the players by their positions are shown in Appendix D.

Players' OFF vs. DEF Ratings: Possessions more than 3000

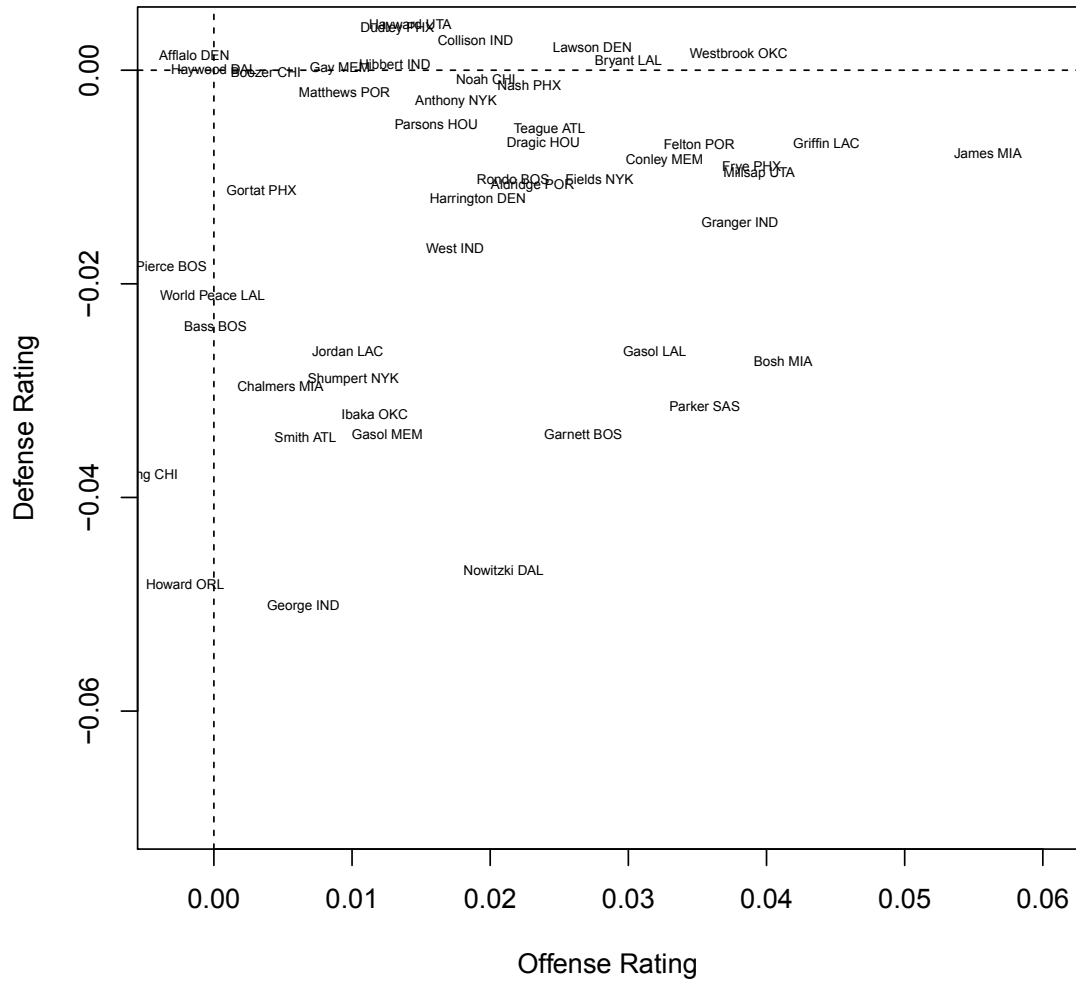


Figure 4.3: Players' Offensive vs. Defensive Ratings: Possessions over 3,000 with + Offensive and - Defensive Ratings

We also plot observed against estimated ratings for players. Similar to the definition in the teams' offensive/defensive observed ratings, players' observed offensive/defensive ratings are the percentage of total number of scoring events observed against total number of offense or defense possessions for each player. The shrinkage effects seem to be influencing many of players to agree with their observed ratings.

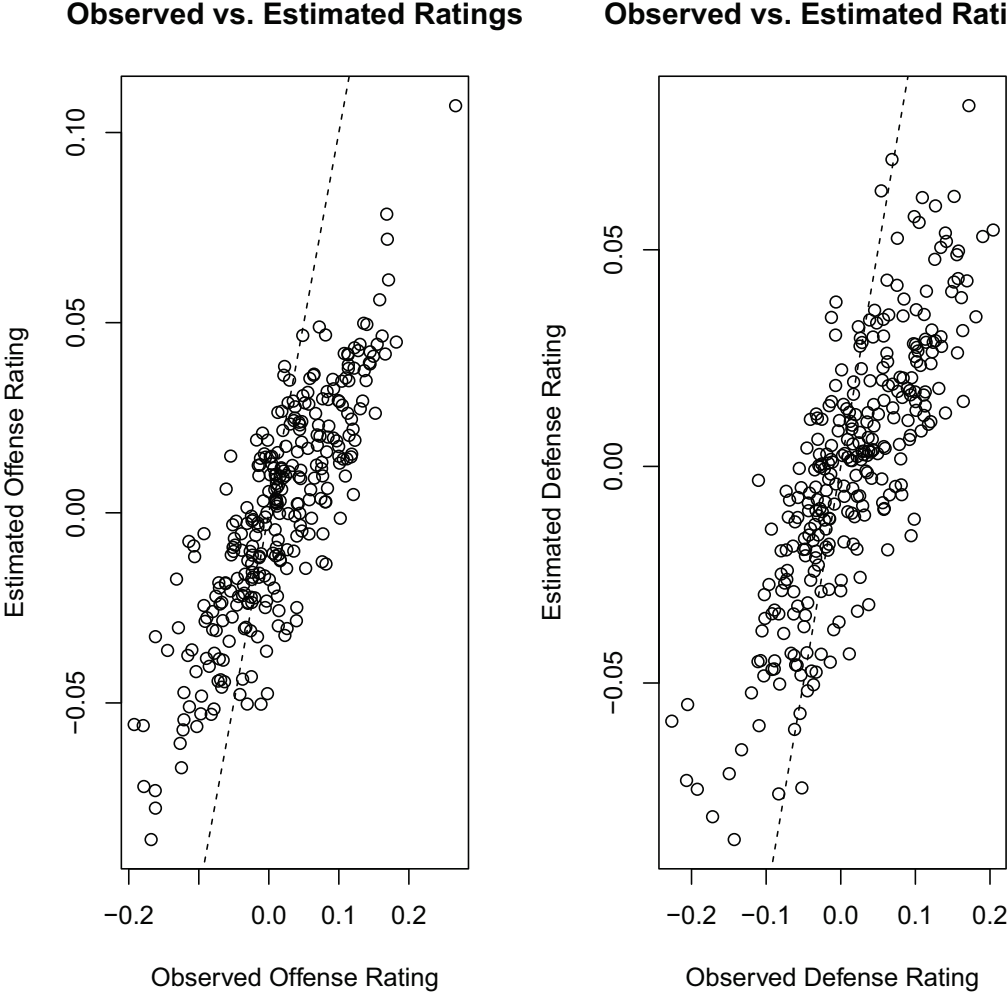


Figure 4.4: Observed and Estimated Players' Offense/Defense Ratings

4.3 Discussion

Our method of using a penalized logistic regression with L2-regularization improves and overcomes limitations of simple and adjusted plus/minus. We show that our estimated offensive/defensive ratings closely match with the NBA teams' or players' observed abilities. We mention that our method models partial effects of the teams' and players' offensive/defensive abilities, rather than marginal effects by plus/minus; this allows us to compare the abilities across teams and players, as opponent team or teammates' abilities are controlled for in the regression. This comparison is more intuitive for our method when it comes to distinguishing offensive and defensive ratings - our offensive/defensive ratings are in the same scale, and this allows intuitive comparison of ratings, of offensive and defensive ratings. However, by changing the response variable calculation for offensive/defensive plus/minus, Ilardi and Barzilai's method does not allow comparisons of two measures. Our method also overcomes limitations of the similar approach by Gramacy et al. (2013) by incorporating shrinkage term, proposed by Thomas et al. (2012), to address the issue with a relatively large number of predictor variables (Gramacy et al., 2013). Through shrinkage, we also overcome the limitation of the adjusted plus/minus model, which have high variance and noise by the outlier players who only have a few, but exceptional plays in the season. Furthermore, our method has an application to unseen game outcome prediction, which is described in the following section.

4.4 Game Outcome Prediction

We estimate each team's and player's offensive and defensive contribution to the outcome of the game. With these effects, we can estimate the probability of a team winning against its opponent team. We recognize that scoring variable S has binomial distribution as the following:

$$S \sim Binom(N, \pi) \tag{4.3}$$

With probability mass function

$$\binom{N}{k} \pi^x (1 - \pi)^{N-x} \tag{4.4}$$

π here presents the probability of a team scoring against its opponent team. Given that team i is playing against team j , i 's probability of scoring in a game is given as the following:

$$\pi = \frac{e^{\omega_i - \delta_j}}{1 + e^{\omega_i - \delta_j}} \tag{4.5}$$

Using the above approach, we simulate the winning probability of the Miami Heat (MIA) win against the Oklahoma City Thunder (OKC), the two finalist in the season's championship. We calculate each team's probability of scoring a basket against each other by the equation in 4.5. With known π , we can simulate S variable for each team in 4.3. The probability of the Heats' win is then calculated by the fraction of simulated scores that are bigger for the Heats. For the tied scores, we give 1/2 weight. We choose $N=100,000$, and number of trials, x , to be 100, which is the average number of possessions of a team in one game.

	MIA	OKC
Scoring Chance	0.54	0.41
Predicted Win	0.59	0.41
Observed Win	0.71	0.29

In 2011-2012 season, the Heats and the Thunders played 7 games - 2 during regular season, and 5 during play-offs. This indicates that 2 games are in-sample, and the other 5 games are out-of-sample. Our results agree with observed outcome of the two teams' matches, favoring the Miami Heat, on the out-of-sample data. Given the small sample size

of the actual matches, we believe that their probability of winning against each other would converge to our results; 7 is not a big enough sample size.

Chapter 5

Conclusion

We introduce penalized regression to improve the popular approach of rating player contributions, adjusted plus/minus. We describe applications and limitations of plus/minus, improvement to adjusted plus/minus by Illardi and Barzilai (2008), and other methods proposed in ice hockey by Thomas et al. (2012) and Gramacy et al. (2013). Incorporating core features of each method, we propose a penalized logistic regression with L2-regularization for estimating offensive/defensive abilities of teams and players separately. This method overcomes limitations of adjusted plus/minus approach, and also addresses limitations of the above mentioned methods. Furthermore, it discovers the underrated players' true abilities. We validate our method by comparing the estimated ratings with observed probabilities of each player scoring, and juxtaposing them with shown performance of the players. We discuss its application to predicting the unseen games' outcomes of the Miami Heat against the Oklahoma City Thunder, and we are able to accurately predict the outcomes.

5.1 Limitations

Despite its effectiveness, we identify a few limitations to our method. The shrinkage value, λ , is not working in the way we expect. As mentioned, below-average, replacement players were showing higher ratings than star players, due to over-shrinkage of star players. This also leads to an issue that the estimated offensive/defensive ratings are not completely aligned with the known abilities of players. Some of the best offensive/defensive players were not among the top ratings as we expect, and average players seemed to get higher ratings than some of those star players. This is due to the fact that we are only looking at the scoring event as our response, and estimating players' probabilities of scoring. Scoring alone cannot quantify the players' offensive and defensive abilities. We identify a possible correlation issue with some players, whose abilities are comparable, as observed in defensive ratings of Jeffries and Chandler in the New York Knicks. We also suspect that that we have a relatively small sample size in this regular season compared with the other regular seasons - there 15,819 observations with 519 independent predictor variables. Our modification of collapsing below-average players into a single entity partially solved the issue. We expect this issue will be solved by incorporating multiple seasons' data.

5.2 Future Work

Our method will serve as a cornerstone for predicting the outcomes of unseen games in basketball. Specifically, we want to expand our approach to ordinal regression, which enables prediction of different scores of basketball: +1, +2 and +3. This will improve predicting power of game prediction, as we can estimate the boxscores close to real basketball matches. Also, we can expand our model to incorporate another metrics that may be relevant in the model. Not to mention that there are increasingly higher stakes for correctly predicting the brackets, as seen in Warren Buffet's recent \$ 1 billion challenge for the perfect National

Collegiate Athletic Association Basketball (NCAA) brackets. We hope to apply our improved method to get better predictions of the basketball basket. Ultimately, we want to develop a ranking system, that improves the current power ranking, to reflect the teams' true abilities that are comparable across other teams, and seasons.

References

1. Beech, R. (2003). *NBA Roland Ratings!*. 82games.com. Retrieved from <http://www.82games.com/rolandratings.htm>
2. Blewis, B. (2012). *2012 NBA All-Average Team*. takingbadschotz.com. Retrieved from <http://www.takingbadschotz.com/?p=6360>
3. Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Journal of Statistical Software, Vol. 33, 1, 1-22.
4. Ilardi, S. and Barzilai, A. (2004). *Adjusted Plus-Minus Ratings: New and Improved for 2007-2008*. 82games.com.
5. Lewin, D. (2007). *2005-2006 Adjusted Plus-Minus Ratings*. 82games.com. Retrieved from <http://www.82games.com/lewin2.htm>
6. Rindner, Grant. (2012). *Thabo Sefolosa and the Next Generation of Superstar Stoppers*. Bleacher Report. Retrieved from <http://bleacherreport.com/articles/1189966-thabo-sefolosa-and-the-next-generation-of-superstar-stoppers>
7. Rosenbaum, D. T. (2004). *Measuring How NBA Players Help Their Teams Win*. 82games.com.

8. Shalizi, C. (2012) *Chapter 12: Logistic Regression*. Retrieved from <http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf>
9. Sporting Charts. *Basketball Plus-Minus*. Retrieved from <http://www.sportingcharts.com/dictionary/nba/basketball-plus-minus.aspx>
10. Thomas, A. C., Ventura, S. L., Jensen, S., and Ma, S. (2012). *Competing Process Hazard Function Models for Player Ratings in Ice Hockey*. Tech. rep., ArXiv:1208.0799.
11. Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society, Series B (Methodology), Vol. 58, 1, 267-288.
12. Witus, E. (2008). *Calculating Adjusted Plus/Minus*. Count The Basket. Retrieved from <http://www.countthebasket.com/blog/2008/06/01/calculating-adjusted-plus-minus/>
13. Zwerling, J. (2012). *Tyson deserving of Defensive Player award*. ESPN. Retrieved from http://espn.go.com/blog/new-york/knicks/post/_/id/15989/tyson-deserving-of-defensive-player-award

Appendix

A. Team Abbreviation

Team Abbreviation			
ATL	Atlanta Hawks	MIL	Milwaukee Bucks
BOS	Boston Celtics	MIN	Minnesota Timberwolves
CHA	Charlotte Bobcats	NJN	New Jersey Nets
CHI	Chicago Bulls	NOH	New Orleans Hornets
CLE	Cleveland Cavaliers	NYK	New York Knicks
DAL	Dallas Mavericks	OKC	Oklahoma City Thunder
DEN	Denver Nuggets	ORL	Orlando Magic
DET	Detroit Pistons	PHI	Philadelphia 76ers
GSW	Golden State Warriors	PHX	Phoenix Suns
HOU	Houston Rockets	POR	Portland Trail Blazers
IND	Indiana Pacers	SAC	Sacramento Kings
LAC	Los Angeles Clippers	SAS	Sacramento Kings
LAL	Los Angeles Lakers	TOR	Toronto Raptors
MEM	Memphis Grizzlies	UTA	Utah Jazz
MIA	Miami Heat	WAS	Washington Wizards

Table 1: Team Abbreviation Code for NBA Teams

B. Cross-Validation for Team Model

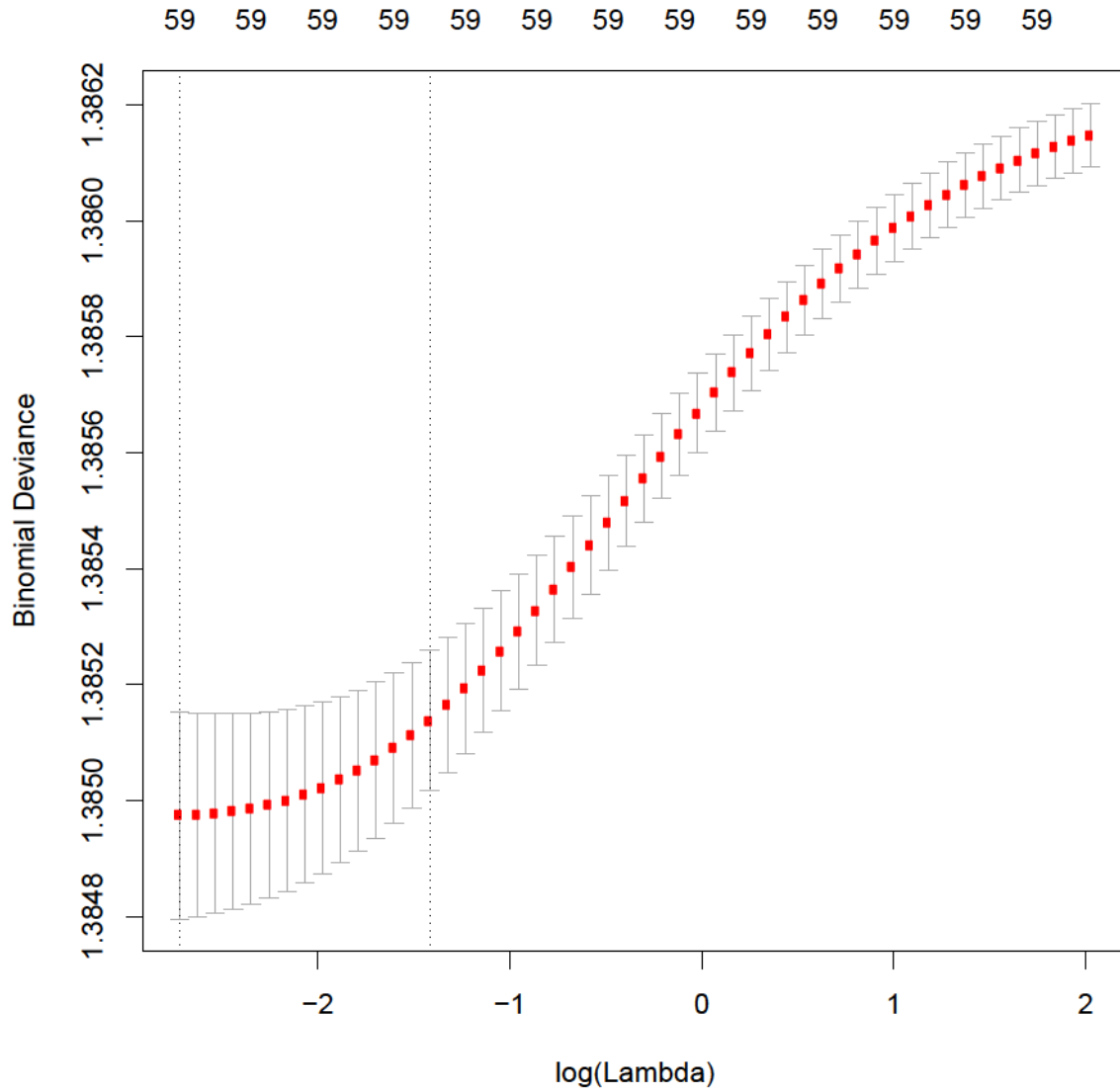


Figure 1: 10f-fold Cross-Validation for shrinkage λ selection for Team Model

C. Cross-Validation for Player Model

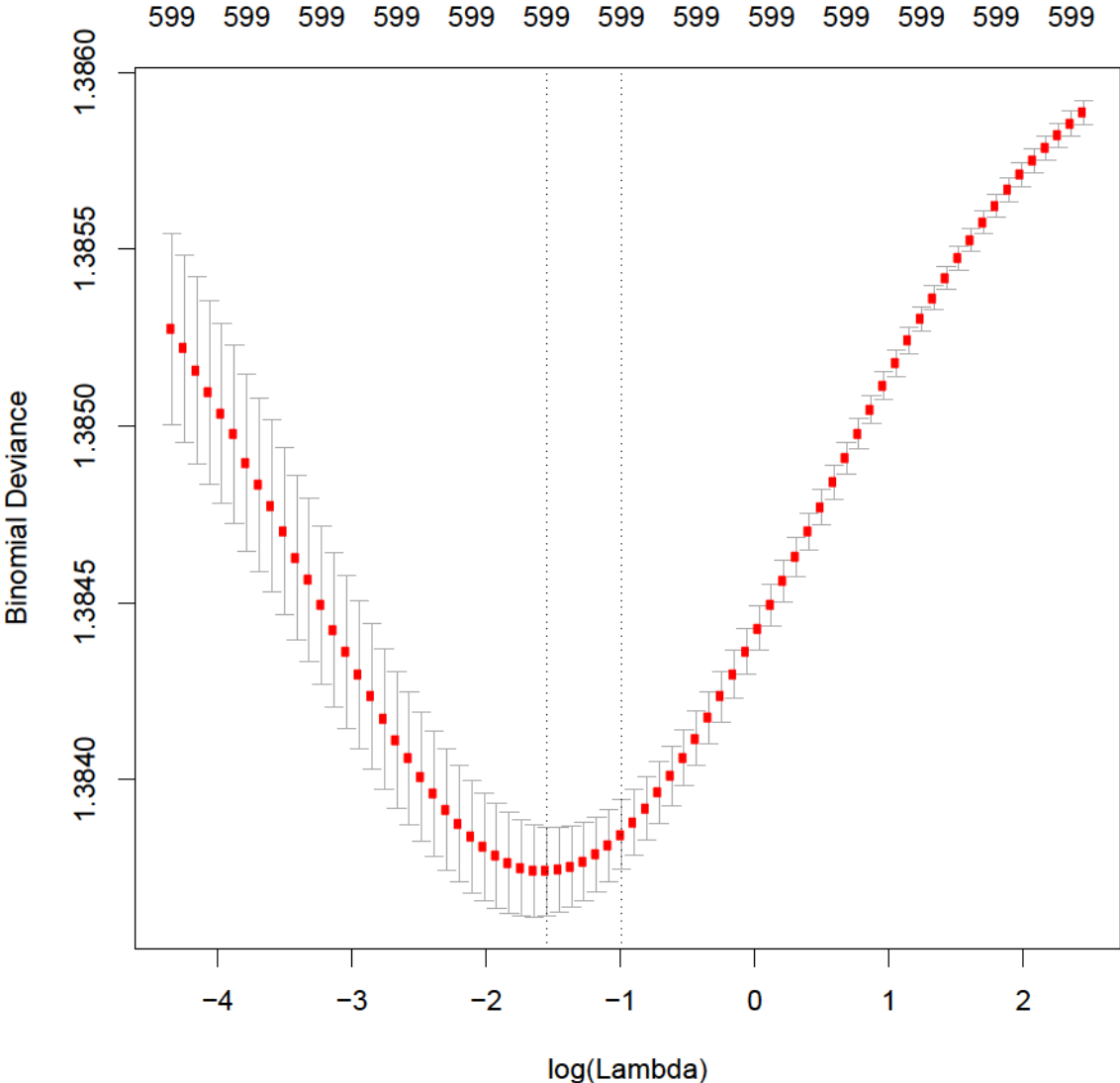


Figure 2: 10f-fold Cross-Validation for shrinkage λ selection for Player Model

D. Players' Offensive/Defensive Ratings by Positions

OFF vs. DEF ratings for players for position: Point Guard

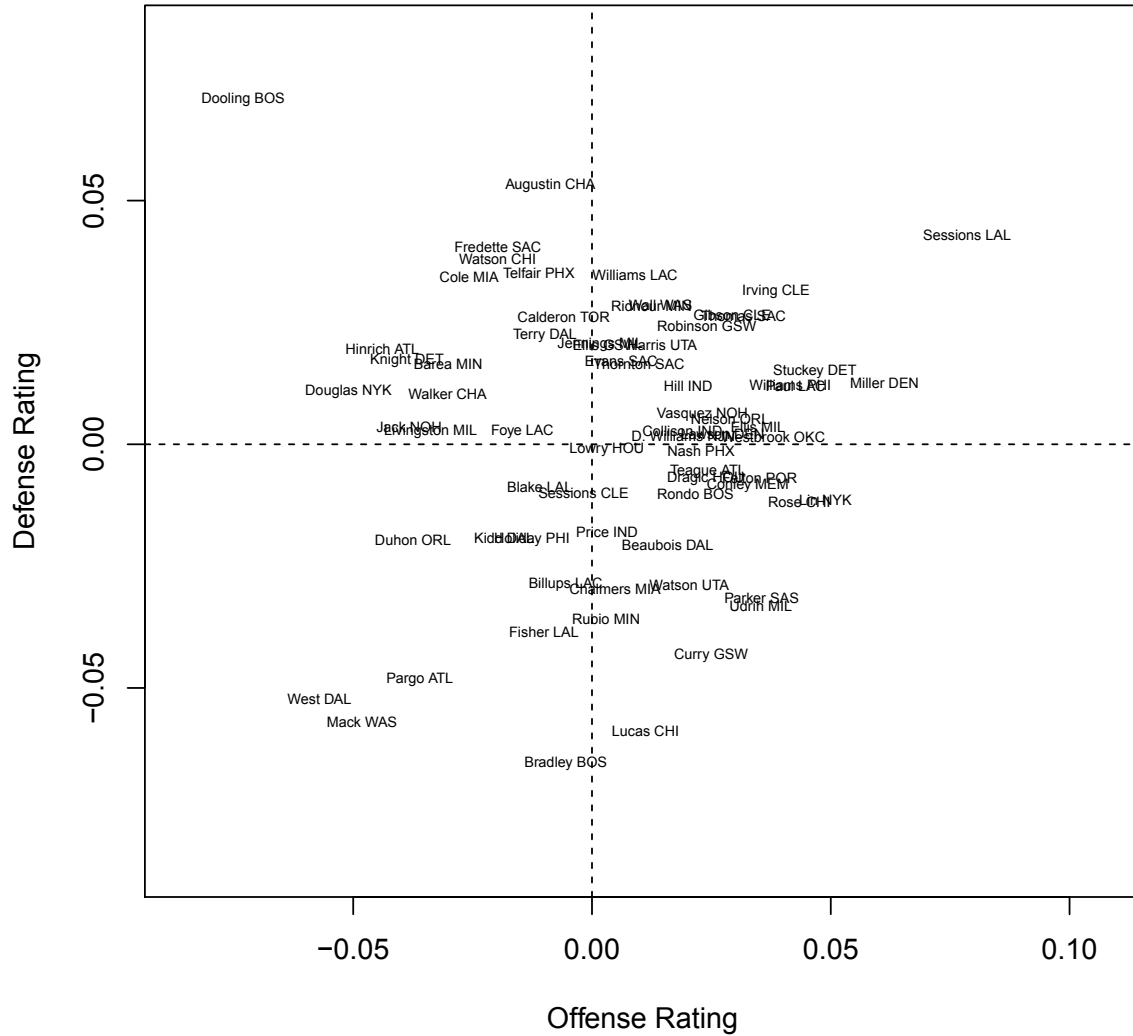


Figure 3: Players' Offensive/Defensive Ratings for Point Guard

OFF vs. DEF ratings for players for position: Shooting Guard

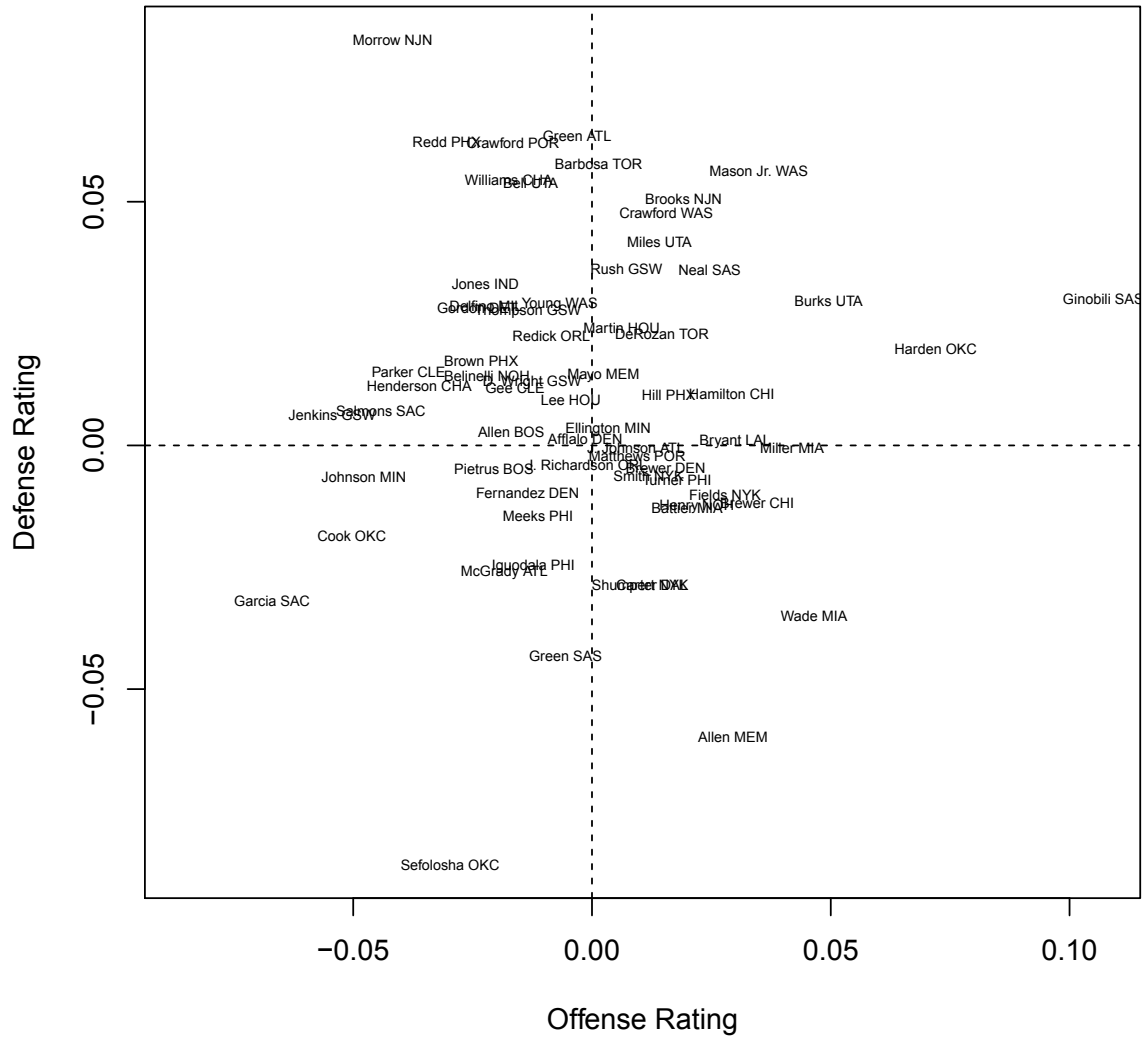


Figure 4: Players' Offensive/Defensive Ratings for Shooting Guard

OFF vs. DEF ratings for players for position: Power Forward

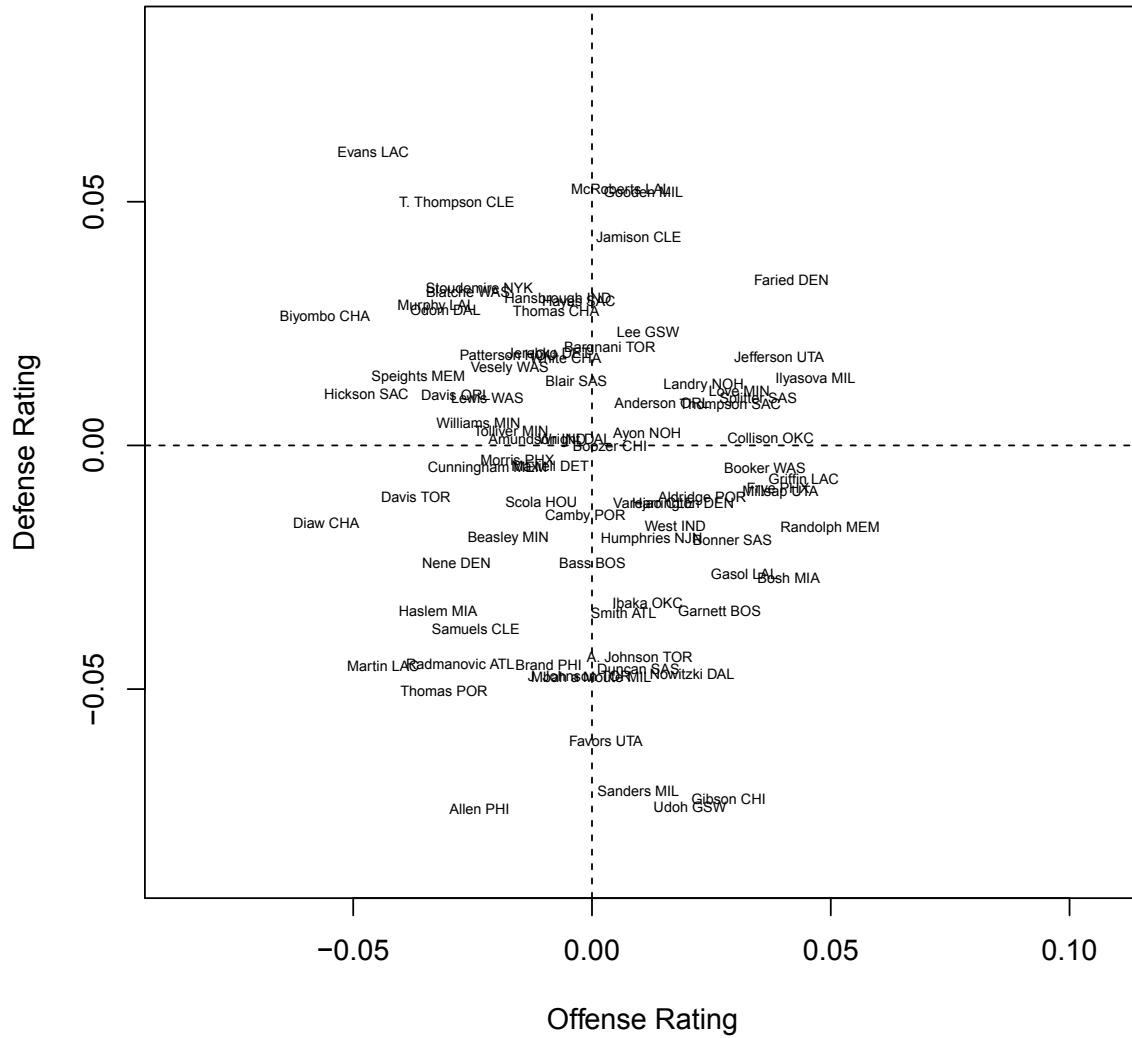


Figure 6: Players' Offensive/Defensive Ratings for Power Forward

OFF vs. DEF ratings for players for position: Center

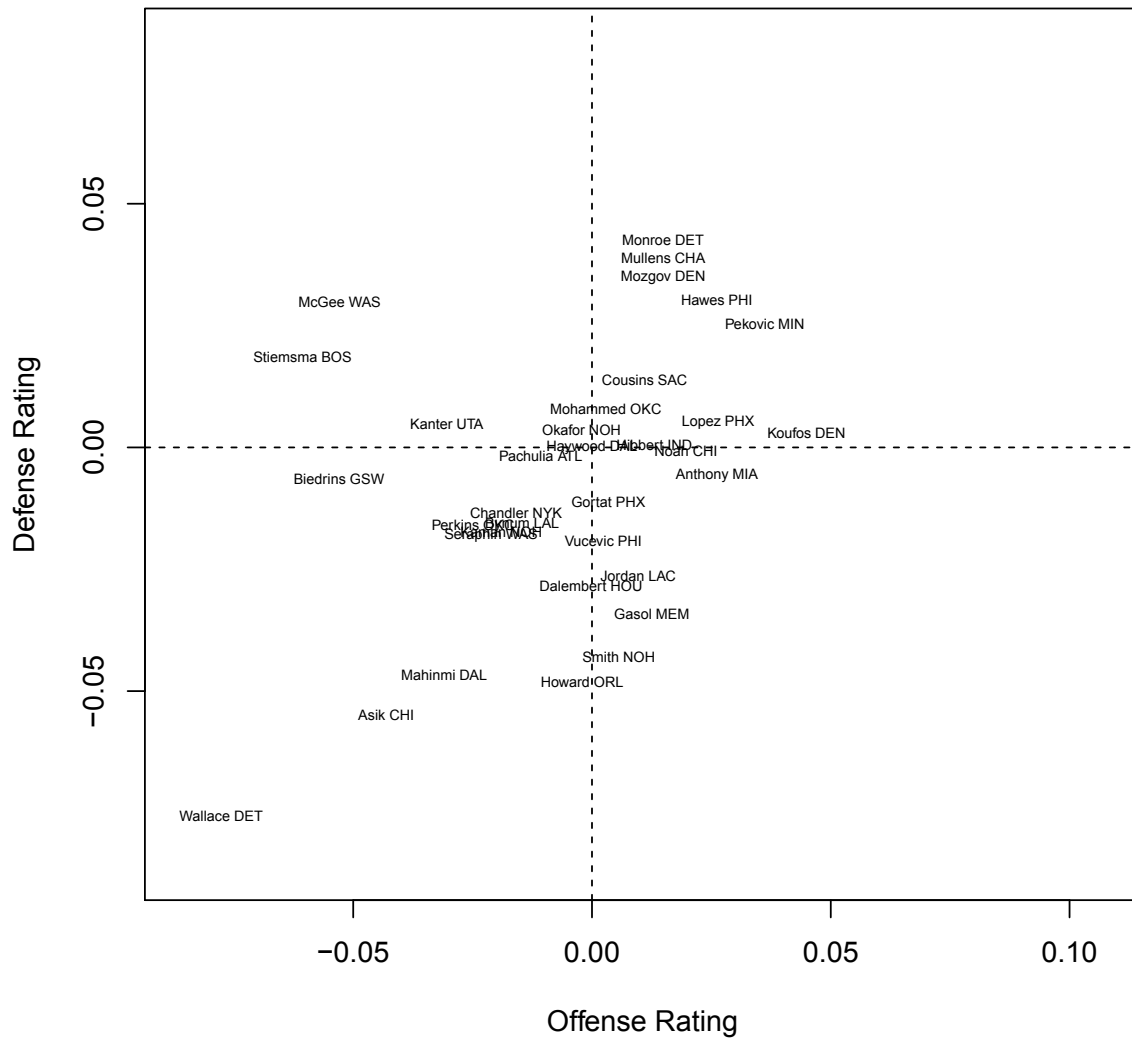


Figure 7: Players' Offensive/Defensive Ratings for Center