A DYNAMIC MODELLING AND OPTIMIZATION

APPROACH TO DAILY FANTASY BASKETBALL



Approved by:

_____

Dr. Jing Cao
Associate Professor of Statistics

_____

Dr. Tom Fomby
Professor of Economics

_____

Dr. Lynne Stokes
Professor of Statistics

_____

Dr. Wayne Woodward
Professor of Statistics

# A DYNAMIC MODELLING AND OPTIMIZATION APPROACH TO DAILY FANTASY BASKETBALL

A Dissertation Prospectus Presented to the Graduate Faculty of the

Dedman College of Humanities and Sciences

Southern Methodist University

in

Partial Fulfilment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Statistical Science

by

Charles South

B.S., Applied Mathematics and Statistical Science, Southern Methodist University, 2006

M.P.S., Applied Statistics, Cornell University, 2007

M.S., Statistical Science, Southern Methodist University, 2014

May 14, 2016

ACKNOWLEDGMENTS

I have been blessed to work with tremendously bright and open-minded faculty members; in particular, I am forever grateful for the guidance, patience, and wisdom from my advisor, Dr. Jing Cao. I could not have completed this work without the support and love from my family and friends, as well as my peers who always challenged me throughout our coursework together and pushed me to improve.

South, Charles M.S., Statistical Science, Southern Methodist University, 2014

A Dynamic Modelling and Optimization Approach to Daily Fantasy Basketball

Advisor: Dr. Jing Cao

Doctor of Philosophy conferred May 14, 2016

Dissertation completed April 30, 2016

Fantasy sports, particularly the daily variety (where a new team is selected each day), is a rapidly growing industry with company valuations of $1 billion for the largest players in the business. This dissertation focuses on the use of statistical procedures to develop a system for analyzing daily fantasy basketball, including both the prediction of player performance and the construction of a team. Frequentist and Bayesian paradigms are explored in a myriad of ways to model an aggregate measure of performance, with the efficacy of each model determined by retrospectively simulating daily results from the 2013-2014 season.

Upon choosing the best model, the predictions are used to construct teams under the constraints of the game, typically related to a fictional salary cap and player positions. A number of methods of construction are considered: a permutation based approach, an integer linear programming based approach, a frequentist and Bayesian logistic regression, quadratic discrimination analysis, support vector machines, random forests, and K-nearest neighbors. The methods are compared in terms of the identification of "successful" teams – those who would be competitive more often than not.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

Participating in sporting events – as a competitor or a spectator – is one of the most popular pastimes across many societies in today's world. Since the early 1980's, however, a new method of participation has emerged: fantasy sports. Fans are able to select players from their favorite teams, compile points based on their players' performances in real time, and "compete" against other fans who have built teams of their own. It gives fans a feeling of being closer to the players and the games they play.

Building off the increasing popularity of fantasy sports, companies have been created that facilitate the betting of money on the performance of fantasy sports teams. According to Forbes, one of the largest companies in this arena has more than 1.1 million customers and brought in over $50 million in revenue in 2014[1]. As the competitive nature of fantasy sports has evolved along with the game itself, so has the predictive analytics. Numerous entities – both professional and amateur – have attempted to build models and predict how players will perform based on their underlying profiles. However, very little research has been published in academia revolving around this topic. This research focuses on modelling player performance in one sport in particular – professional basketball. The combination of a reasonable sample size each season (NBA teams play 82 games, compared to just 16 in the NFL) along with a reasonable amount of data generated during each game (as opposed to baseball, where most batters only get 3-4



[1] http://www.forbes.com/sites/stevenbertoni/2015/01/05/how-fanduel-is-turning-fantasy-sports-into-real-money/

opportunities per game to compile statistics) allow for more flexible analyses. We begin by building a wide range of statistical models to predict an aggregate measure of performance for a single player in a single game, which is a linear combination of traditional counting statistics. Next, we expand on the methodology to predict the traditional counting statistics themselves and *then* combine them to form a prediction for the aggregate measure of performance. Finally, the predictions are used to construct "lineups" of players with respect to the predicted aggregate measure and the constraints imposed by the fantasy game itself (Section 1.2 details the rules of the game). A wide range of statistical methods are used to explore the best route to identifying lineups that are likely to be successful. The end result is a completely automated system that pulls data from the internet, builds and implements a model, then uses the results to identify lineups that can be played in a daily fantasy game.



## 1.1 Fantasy Sports Background

Fantasy sports leagues can be broken down into two major categories: season long and daily. Season long fantasy sports leagues are typically composed of 6-14 managers who select their teams via a draft before the season begins. Each player can only be "owned" by one manager, so the draft involves quite a bit of strategy. After the draft, managers can make trades with other managers in his/her league, add new players that were not drafted, and drop existing players. They manage their teams throughout the entire season and accrue points based on the scoring system chosen for that particular league, with the two most common variants being head-to-head scoring and rotisserie scoring. In head-to-head leagues, managers compete directly with one other manager in several statistics (such as home runs, batting average, and earned run average in baseball, among others). At the end of the matchup period (typically a week),

managers earn 1 point for each statistic that exceeds his/her opponents, with the final tally depending on the number of statistics. This cycle is repeated throughout the season with the matchups permuted among all managers, and at the end of the season the manager with the best record is the champion. In rotisserie scoring, teams are ranked from first to last in each of several statistical categories. Each fantasy team receives points in each category based on how they rank in the league, and the overall ranking for a team is typically the sum of all the individual ranks. The season-long fantasy game can be thought of as a marathon – managers must pick players who they think will perform consistently all season, will avoid injuries, and who compliment the other players on his/her team.

The other variant of fantasy sports games – analogous to a sprint – are daily fantasy leagues. These leagues can consist of anywhere from two managers to hundreds or even thousands of managers. Managers can enter leagues with one team or multiple teams, depending on the setting. Regardless of the format, managers are given a fictional salary cap and all players are given a fictional salary tied to previous performance. The strategy involves picking the best players who also offer the best value relative to their salary for that particular day. The same player can be owned by multiple managers in this setting, so in theory there can be duplicate lineups. Most mainstream scoring systems use a linear combination of counting statistics as the measure of player performance, meaning each player will have one "score" at the end of the day. The manager with the highest sum score of all his/her players is the winner of that league. It is this variant of fantasy sports that has exploded in popularity over the last few years and which offers a challenging statistical question: what factors affect player performance on a given night, and how much of the variability in performance can be accounted for using statistical models?

*1.1.1 Introduction to Basketball*

The game of basketball is one of the most recognizable and popular sports around the world. It is a team sport in which teams of five compete against each other (with substitute players available), trying to shoot a ball through a hoop. The five positions on each team are the point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C). The guards are usually the smallest players on the court, followed by the forwards. The center is traditionally the largest player. The point guard is an important position because this player has the ball in his hands most often while on offense and tries to facilitate easy opportunities for his team to score points by running set plays or making good passes. Points can be scored one, two, or three at a time by making a free throw, a two point field goal, or a three point field goal, respectively, subject to a large number of rules and regulations that have been developed over the years. In the National Basketball Association (NBA), there are four quarters that span 12 minutes each and the team with more points at the end of the fourth quarter is declared the winner. In the event of a tie, the opposing teams will play a short overtime period to determine a winner. All teams in the NBA play a total of 82 games a season.



One of the main benefits to analyzing daily NBA player performance is the sheer volume of data available. While Major League Baseball (MLB) data initially garnered more attention in the analytics world due to its label as “America’s pastime” and the statistical simplicity of every play (there is only a single pitcher throwing to a single batter), most batters only get 4 attempts to hit in a single game and most starting pitchers only pitch once a week. A 162 game schedule means that performance is reasonably predictable over the course of a season, but not on a day to day basis. In the NFL, teams only play 16 games a season and there are a total of 22 players on the field at a time, making it difficult to separate and predict individual performances. On the other hand, in the NBA the best players play 30 or more of the 48 potential minutes in each game

and accumulate many different types of statistics (a full glossary of metrics available in this research is included in the Chapter 2 Appendix). Teammate interaction is important in the NBA, but the number of metrics that can be captured for an individual player allow for reasonably reliable predictions. For this reason, basketball was chosen as the focal point of the research.

### *1.1.2 Daily Fantasy Basketball Rules*

Recall, the spirit of fantasy sports is to allow fans to pick a team of real players and compete using the statistics they accrue during games. While there are many websites that host daily fantasy games, the main rules of the game remain the same: under the constraint of a fictional salary cap, pick a set of players who accrue points based on some aggregate measure of observable statistics; the main differences are in the fictional dollar amounts (and, as a result, player costs) and the weights used to calculate the player's score.

The website whose rules this research utilized is DraftKings, arguably one of the two biggest daily fantasy game hosts in the market. Their measure of a player score – call it "fantasy points" (FP) – is defined as

$$FP = X_1 + 1.25X_2 + 0.5X_3 + 1.5X_4 + 2X_5 + 2X_6 - 0.5X_7,$$

where the variables are as follows:

$X_1$ = total number of points scored by the player

$X_2$ = total number of rebounds (missed shots) grabbed by the player

$X_3$ = total number of three point field goals made by the player

$X_4$ = total number of assists (a pass to a teammate who immediately scores) made by the player

$X_5$ = total number of steals (legal takeaways from the opposing team) by the player

$X_6$ = total number of blocks (shot attempts by the opposing team swatted away) by the player

$X_7$ = total number of turnovers (giveaways to the other team) by the player

The FP metric is designed to be a representation of the overall performance by a player on a given day, with the goal of the daily fantasy game being to accrue a team of NBA players with the largest FP sum for the day. However, the team must be constructed in a way that mirrors a real team – one each of the traditional five players must be selected (point guard, shooting guard, small forward, power forward, center), as well as an additional guard (either point or shooting), forward (either small or power), and utility (any of the five positions). The team of players must be selected from NBA teams actually playing on the particular day of interest; the NBA players do not, however, have to be from the same team. The fan choosing the team then decides on the type of daily fantasy game to play as well as the amount of money he/she would like to wager (games range from free to costing over $10,000 to join). At the end of the day, the sum of the FP for his/her eight players is compared to all other FP sums of fans that entered the same game and a winner is declared.



## 1.2 Literature Review

With the daily fantasy sports phenomena reaching peak popularity only recently, little academic literature exists when it comes to predicting daily player performance. Casals and Martinez (2013) used mixed models with random effects to predict both points scored and win score (a linear combination of counting statistics similar to FP that is a different indicator of player performance). They found that the player, difference in team quality, age, player position,

interaction between age and player position, minutes played, and usage percentage (an estimate of the percentage of team plays used by a player while he was on the floor) were significant when predicting win score. The player, difference in team quality, whether the player started, age, player position, minutes played, and usage percentage were significant when predicting points. However, they used a filtering process to create a balanced study design with repeated measures, resulting in only 27 players (who played 81 games) in their study. Further, they only fit a single model using all the data at once rather than fitting daily models and tracking performance. Lastly, they did not consider a wide range of countable statistics as potential explanatory variables.

There are numerous websites that offer specific daily predictions, but they are for-profit and do not reveal any of their methodology. Nonetheless, there have been attempts at quantifying player ability in a more generic way. Kubatko *et al.* (2007) introduced a wealth of potential advanced metrics that can be used in an analysis, with a central focus on statistics related to possessions (a possession starts when one team gains control of the basketball and ends when they give it up, either after a score or a turnover). While informative, they did not carry out any specific analyses related to player performance. Page *et al.* (2013) used a hierarchical Bayes model with Gaussian process regressions to estimate individual player production curves based on player position, average number of minutes played, and usage rate. However, the principal goal of their research was to estimate an entire career production curve rather than game-to-game results. With a similar goal in mind, Page and Quintana (2015) introduced a sophisticated method of modelling career trajectories using Bayesian penalized B-splines and clustered them based on curve-smoothness and subject-specific covariates (age, experience, and draft order). The purpose of this research, much like Page's previous paper, was to differentiate trends in

performance over the course of a player's career with the goal of helping management make personnel decisions rather than focusing on game-by-game results to identify players who are likely to perform well on a given night.

Fearnhead and Taylor (2011) used Bayesian methods based on play-by-play data as well as box score data (aggregate results for a single game) to give a single measure of latent offensive and defensive ability. They used their framework to estimate the posterior probability that one player is stronger than the other. Piette, Pham, and Anand (2011) used network analysis and bootstrap testing of play-by-play data to estimate how important a player was relative to his five man unit, as well as how well he performed in that role statistically. In particular, they used eigenvector centrality with a random restart to determine the importance of a player in a network and ranked the players based on his centrality score. However, their algorithm was based only on points as a unit of measure and resulted in a latent metric. Arkes and Martinez (2011) showed that the more success the home team had in their previous five games the more likely they were to win. Entine and Small (2008) found that the well-known home court advantage NBA teams have is partially explained by the tendency of schedules for the travelling team to have reduced rest. The results from these final two papers do not provide much insight for player-level analyses when considering that an already small team advantage was created by a whole group of players. It should be noted that none of the papers in our literature review address day-to-day prediction of player performance.



### 1.3 Research Overview

As both statisticians and fans of basketball, we were excited to have the opportunity to address a topic with such limited academic exposure. The objective of the research was to

provide a systematic approach to daily fantasy basketball; that is, develop an automated and integrated system that pulls the data, runs the statistical model to make predictions, generates a pool of potential lineups using the predictions, and then returns a number of lineups (e.g. the top 10) that are most likely to be successful. Two different dynamic modelling approaches are explored using box score data from the 2013-2014 NBA season: one using FP as the response variable and the other building seven separate models for the statistics that compose FP. Within these major paradigms, a wide variety of models are built and tested in both the frequentist and Bayesian frameworks using R and WinBUGS. We demonstrate empirically that a Bayesian model with player specific intercepts, slopes, and variances is the optimum model in the first modelling approach, while a similar Bayesian model with a mixture component to account for inflated 0 values is optimal in the second approach. Next, using data from the 2014-2015 season, we proceed to the idea of choosing optimal lineups by exploring the performance of a wide variety of methods. After generating potential lineups using a simple permutation approach and an optimized routine using an integer linear programming model, we consider predicting the likelihood that a lineup will exceed a pre-determined FP threshold by using logistic regression in both frequentist and Bayesian frameworks, quadratic discriminant analysis, support vector machines, random forests, and *k*-nearest neighbors. The top performing classification methods are then implemented in a larger sample of data from the 2015-2016 season, and we illustrate that, in general, the nonparametric methods outperform the parametric methods, with the best chance of being exposed to successful lineups coming from *k*-nearest neighbors with $K = 75$.