Spring 2021

Does Defense Actually Win Championships? Using Statistics to Examine One of the Greatest Stereotypes in Sports

Thomas Burkett
University of South Carolina - Columbia, tdb@email.sc.edu

DOES DEFENSE ACTUALLY WIN CHAMPIONSHIPS? USING STATISTICS TO EXAMINE ONE OF THE GREATEST STEREOTYPES IN SPORTS

By

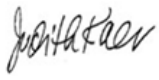
Thomas Burkett

Submitted in Partial Fulfillment
of the Requirements for
Graduation with Honors from the
South Carolina Honors College

May 2021

Approved:

Dr. David Hitchcock
Director of Thesis



Dr. Judith Kalb
Second Reader

Steve Lynn, Dean
For South Carolina Honors College

TABLE OF CONTENTS

THESIS SUMMARY

For years, defense in sports has been hailed as one of the most valuable traits a team or player can possess. This infatuation with defense is most embodied in the classic saying that "defense wins championships." The phrase has its origins in American football but has expanded into other team sports, as well. Despite being a relatively old saying, little research has been done in recent sports history to test whether a team's defense can carry it to a championship win.

The past decade of play in the National Basketball Association (NBA) has seen a rise in both 3-pointer attempts and overall scoring averages per game. The result is a league centered around offensive analytics and efficiency. More than ever, teams prioritize offensive productivity over defensive fundamentals. This study collected data from the past 10 NBA regular seasons and tested whether one of sport's oldest sayings held true for a modern NBA. Defensive performances were measured with an advanced defensive metric, a team's defensive rating, and other defensive statistics based on per game averages. The mean defensive ratings of championship teams and non-championship teams were compared through statistical analysis to determine if defensive performances truly differed between these groups. Mathematical regression models were then composed to analyze how well championship teams were predicted from offensive and defensive statistics. The different models were compared against each other based on how well they fit the data set, which statistics they found to be predictors of championship teams, and how well they predicted potentially new observations. The same statistical tests and models were then used for two other populations: teams that were one of the last four teams in their respective postseason and those that were not. This established a distinction between defense being predictive of winning a championship or only putting teams in a situation where they were close to winning a championship. The goal of this study was to

examine the importance of defense in winning an NBA championship today and whether similar results could be achieved from offense alone and a combination of offense and defense.

The results found defense and defensive statistics to be significant predictors of championship teams and teams that were close to winning a championship in today's NBA, as well as these teams had better defensive ratings than teams that did not fit this criteria. However, similar results were achieved with offensive statistics and more significant results were obtained from a combination of offensive and defensive statistics. Excellent defensive performances were predictive of championship teams, but offensive context showed that defense on its own does not win the most important games of the season. The past 10 NBA champions were consistent in dominating on both the offensive and defensive sides of the basketball. "Defense wins championships" holds some truth in the modern NBA, but fails to acknowledge that balanced teams have been the true winners of NBA championships.

INTRODUCTION

One of the most frequently stated axioms in sports is “defense wins championships.” The idea is intuitively sound; a strong defense prevents the opposition from scoring large amounts of points and therefore increases a defensive team’s chances of winning. Teams in the National Football League (NFL) have shown that a solid defense is as equally important as a solid offense in terms of predicting team success (Robst, VanGlider, Berri, and Vance, 2011). However, in a fast-paced game such as basketball, all players must be able to play on both sides of the ball and defense from possession to possession is more crucial than in football. Strong offenses may be able to produce points, but they are unable to cover for a poor defense by dominating time of possession due to shot clock restrictions. Similarly, strong defensive teams must be able not only to stop the opposition but to put up points of their own.

Defense in basketball has been examined on many levels. College Division I men’s basketball teams with above average defensive statistics have been found to have better win-loss records than teams with average statistics (Mondello, 2000). A team’s record is highly important in the NBA, as the teams that finish the season with the most wins enter the playoffs with higher seeding and therefore gain an advantage over lower seeded teams by having four out of a possible seven games in a playoff series played on their homecourt. Mondello also noted that a team’s schedule is beyond its control, and this could lead to talent differentials (2000). A commonly used metric in the NBA is a team’s Strength of Schedule (SoS) to assess the relative strength of competition a team faces throughout the course of the regular season, and this metric varies depending on the conference and division a team plays within (Cappe, 2020). Defensive statistics are typically reliable measures of a team’s overall defensive performance, but they could be slightly misleading if the opponents a team is facing are especially weak. As valuable as

a winning record may be, NBA teams still have to battle their way through the playoffs against more skilled opponents and a more complex postseason format than that of college basketball. Even having the best record in the NBA does not guarantee a championship winning team, so top defenses might not be predictive of winning the games that matter the most.

Defensive statistics in basketball have also been previously analyzed at an international level. In the Beijing 2008 Summer Olympics men's basketball tournament, the winning teams were disciplined on defense, preventing the opposition from scoring or being fouled on 52% of their defensive possessions (Álvarez, Ortega, Gómez, and Salado, 2009). Due to the Olympic setting, the games in this tournament can be equated to a playoff scenario where teams are playing their hardest to win the championship, or a gold medal. Analysis indicated a significant relationship between man-to-man defenses (a one-on-one defensive scheme where each player on a team guards one other player on the court) and winning, defensive efficacy and winning, and a negative association between allowing inside passes and winning (Álvarez, Ortega, Gómez, and Salado, 2009). Defense clearly has an impact on a game-to-game basis, but the competition in the NBA is more talented than the Olympic level. The talent differential in Olympic play is displayed by the United States men's basketball team's continued dominance in the majority of the Olympics since the inception of the event (*USA Men's National Teams*). Many of the players on the modern United States team play in the NBA, effectively establishing the NBA as the premier basketball league. The best foreign basketball players, who lead their teams in the Olympics, also play in the NBA. The NBA consists of more than 400 of the best basketball players in the world, resulting in the 30 best teams in the world. The offenses are naturally going to be more skilled, making defense as important as it is difficult. The question, then, is how successful the top defensive teams in the NBA are in terms of championship wins.

Great defensive teams are identified by above average defensive statistics. Not all defensive statistics are equally associated with winning, however. Studies of the Spanish Basketball League have shown blocks, steals, and defensive rebounding numbers as indicative of a positive or negative game result (Ibáñez, Sampaio, Feu, Lorenzo, Gómez, and Ortega, 2009). Unlike advanced defensive metrics that have been composed in recent years, these statistics are taken directly from box scores of games and have their impact directly felt on the result. Other statistics that could illustrate defensive impact are a team's personal fouls per game, opposition field goals and attempts, and opposition turnovers per game. Personal fouls can occur on either offense or defense, but they typically are committed by a defense in their efforts to prevent the opposition scoring. Opposing 2 and 3-point field goals, as well as free throw attempts, can be telling of how well a defense is guarding shots, but they are potentially misleading in that the best players in the world can still make heavily contested shots. An opposition's turnovers per game will be strongly correlated with steals per game, though a turnover may still occur from solid defense if the opposition is forced to lose control of the ball out of bounds of the playing area. Most of these statistics have yet to be examined in the context of winning championships in the NBA, opening the door for further statistical exploration.

Advanced metrics still are worthy criteria for evaluation of a team's defensive prowess. Basketball analytics has given rise to statistics called offensive and defensive ratings, which are essentially the amount of points a team scores in 100 possessions or the amount of points the team allows in 100 possessions, respectively (Zuccolotto and Manisera, 2020). NBA teams want to have a high offensive rating, but a low defensive rating. Defensive rating does not tell the whole story of a team's defensive performance, but provides an overall idea of where the team

stands in this regard. If defense does truly win championships, then one would expect NBA champions to have higher defensive ratings than non-champions.

From 1950 to 2014, there existed an almost perfect balance between offense and defense in predicting an NBA team's postseason success; defensive statistics and schemes on their own did not predict postseason success any better than offensive metrics (Otten and Miller, 2015). These findings do not refute the theory that defense wins championships, but rather assert that teams must be equally strong on both sides of the court to have the greatest chance of getting to and winning the big game. Research on player acquisition in the NFL has supported this by finding that there was no significant difference in investing in defense more than offense and noting that a balance of both was most likely to predict championships (Robst, VanGlider, Berri, and Vance, 2011). Sports teams depend on both offensive and defensive talent to be difference makers; an unbalanced team will have its shortcomings exposed regardless of how exceptional its specialty is. Historically, defense may have indeed won championships with the assistance of an equally talented offense.

However, the NBA today is not the same as researched in the past. Over the past decade, the average points scored per game in each NBA season has been steadily increasing (*NBA League Averages – Per Game*). Teams are well aware of this trend and are in fact purposefully trying to score more points than ever by prioritizing 3-point field goals over 2-point field goals. The 3-point shot is more efficient than a midrange 2-point shot, scoring more points per attempt on average. This rewards teams for taking more attempts beyond the 3-point line while eliminating shots several feet inside the line (Shea, 2018). Data from the past 10 NBA seasons confirms that as the average points scored per game has increased, the average 3-point field goals attempted per game has increased as well (Figure 1).

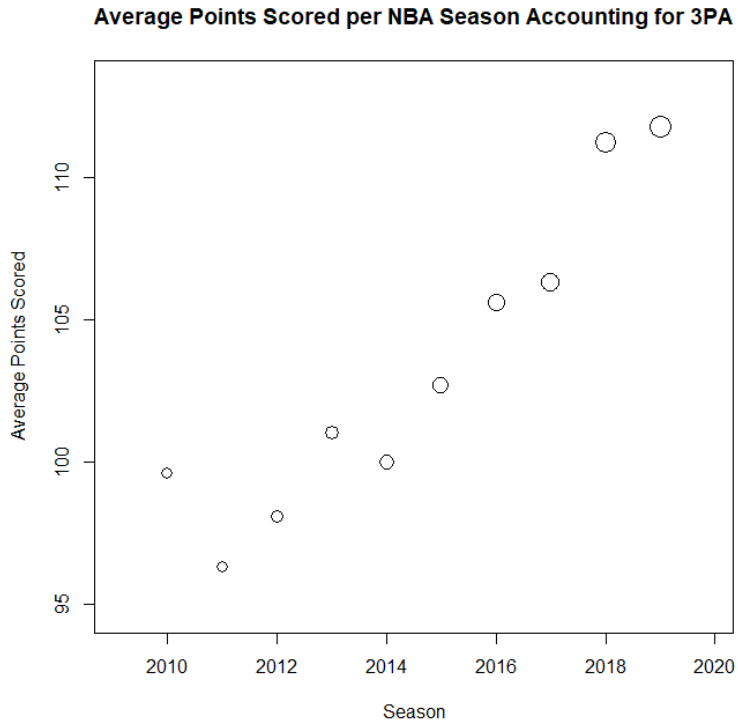


Figure 1. Average points scored per game over the past 10 NBA seasons with 3-point field goal attempts represented by bubble size

More than ever, teams such as the Houston Rockets are incorporating analytics into their playstyle by emphasizing 3-point shooting and efficient offense. Extraordinary defensive teams, like the 2003-2004 Detroit Pistons, are seemingly a trend of the past as the basketball balance begins shifting more and more towards offense. Despite this, the 2020 NBA Champion Los Angeles Lakers were well-known for their defensive prowess and performance during the postseason. Are the Lakers simply an outlier or do top defensive teams still dominate championships in the modern, offensively-favored NBA?

Through statistical modeling and methods, this study plans to determine whether or not the sports stereotype that “defense wins championships” holds merit over the past decade in the NBA and whether defensive statistics are reliable predictors of postseason success on the most competitive basketball stage in the world. After examining whether top defensive teams have

been winning championships in the modern NBA, this study will then predict whether top defensive teams are able to reach the championship of their respective conference in the postseason. A comparison of conclusions from these two models will distinguish whether superb defense is indicative of winning championships or only deep postseason runs. At the end of its quantitative analysis, this study will further comment on the shifting offensive-defensive balance in the NBA that changes every season.

BACKGROUND

The NBA is constructed in a way where each of the 30 teams plays 82 games per regular season to determine their seeding, or positioning, going into the postseason. Playoff seeding ranges from one to eight and each of the Eastern and Western conferences in the NBA has its own seedings. Seeding depends on a team's win-loss record in the regular season; the number one seeds will have the best winning percentages in their respective conferences while the eight seeds will have the worst winning percentages out of playoff teams in their respective conferences. In the context of the NBA, a "higher seed" is a team seeded at a lower number.

The benefits of a team's higher seed are manifested in the opponents faced in each playoff series. Higher seeded teams will face worse opponents record-wise in the first round of the playoffs. The one seed will play the eight seed, the two seed plays the seven seed, and so on. Teams are initially rewarded for their regular season performance by playing teams that, on paper, are worse than themselves, but some teams may play better in the playoffs or may have had a worse record because they were missing key players for many games in the regular season. The quality of competition advantage is not guaranteed, but typically true. As teams advance through the playoffs, they might face teams that are higher seeded (unless they are the one seed) or they may continue to play lower seeded teams (unless they are the eight seed) depending on the results of other series. Higher seeded teams within each playoff series gain a homecourt advantage. To become NBA champions, a team must play and win four best-of-seven-games series. The higher seeded team in each series is given homecourt advantage, meaning they will play four out of a possible seven games in their own arena. One of these four home games is the possible series-deciding game seven. NBA teams have shown to have greater winning percentages on their homecourt, so this is a significant competitive advantage (Kotecki, 2014).

While every playoff team has a chance to win the championship, the regular season rewards those who performed the best with these advantages.

The NBA playoffs are a split knock-out bracket with the Eastern conference teams on one side and the Western conference teams on the other. Therefore, teams will not face an out-of-conference opponent unless they reach the NBA Finals, the fourth and final series that a team must win to be crowned champions. The series before the NBA Finals, the third potential series, is the Conference Championship series, named because it is between the remaining two teams of each conference. Winning the Conference Championship primarily serves to gain entry into the NBA Finals, but teams also receive a trophy and banner from winning this series. The Conference Championship is the second most prestigious award a team can win in the playoffs and participation is still considered being in a championship situation.

The phrase “defense wins championships” has been commonly attributed to former Alabama Football head coach Paul Bryant (Foxworth, 2018). Bryant originally coined this adage in regard to football, but the use has developed over time to cover all of team sports as a whole. Variations have been made to Bryant’s original saying, such as adding that “offense wins games, defense wins championships,” but the promise of a superior defense winning the most important games has remained at the heart of the phrase (Nweiyue, 2020). In spite of how long this saying has been repeated, the sports media has recently questioned the legitimacy of such a theme in today’s professional leagues (Foxworth, 2018). The main aspect missing from these arguments was in-depth statistical analysis and whether numbers are able to back up Bryant’s claims.

LITERATURE REVIEW

Historically, sports literature has only briefly examined the connection between defense and winning championships. Such literature has primarily analyzed this trend in the NFL and football in general. Research about defense winning championships in basketball is sparse and has not to this point covered what has been defined here as the "modern NBA." Defensive statistics, especially in basketball, have only been lightly investigated and have shown significance in predicting winning games. These studies have provided the groundwork for the analyses done in this thesis.

One of the earliest academic works in sports was Mondello (2000). Dr. Mondello, a former coach at the collegiate level, built off the work of previous studies supporting the notion that strong defense does in fact win championships in basketball and mainly targets coaches as the audience. Mondello also noted that talent and scheduling are variables that cannot always be controlled by a team, and the relative randomness of these variables can lead to variation in defensive performance. The study examined data from 315 Division I men's teams in the 1998-1999 season and claimed that defensive field goal percentage was the most significant predictor of whether a team would win a game. The results concluded that better defensive statistics lead to better win-loss records and that these conclusions likely extend to the professional level and women's basketball. It is important to note this study only looked at win-loss record, and not whether the highly rated defensive teams went on to win the national championship.

Ibáñez et al. (2008) further developed the literature of defensive basketball. The goal of the study was to identify which statistics and metrics were most likely to be indicative of a team's win rate and season success. The researchers emphasized season success rather than game-to-game success, which reflects the same mindset as this study. Since the researchers are

from Spain, the data came from the Spanish Basketball League from two different seasons in the early 2000s. Both offensive and defensive statistics were considered. Many offensive statistics were shown to influence season-long success, but discriminant analysis also showed significant effects of blocks, steals, rebounding, and defensive preparation being the difference between winning and losing teams. These statistics may have a significant impact regarding championship success for teams in today's NBA.

Álvarez et al. (2009) also examined basketball on a more foreign stage. This in-depth academic study briefly described past studies that discussed the importance of defensive rebounding and defensive systems to winning games. Álvarez et al. (2009) examined all these variables and others on a game-to-game basis in tournament play. Data were collected from almost every defensive possession in the 2008 Olympics men's basketball elimination rounds in Beijing, extending the notion of championship defense to the international level. Analysis found a significant relationship between man-to-man defenses and winning, defensive efficacy and winning, and allowing inside passes and winning. The study concluded with the finding that winning teams did not foul or allow points on 52% of defensive possessions. While these results support defense winning championships on the highest level of competition, it is important to note that there is typically more difficult competition in the NBA than in Olympic play.

Robst et al. (2011) further developed sports literature in general by focusing on defense in the NFL. The authors explored this topic on the financial level as well as the competitive level. Their article explores whether defense or offense was more important for championship winning teams in the NFL from 1966 to 2009 and how salary caps restrict teams from investing massive financial resources in either side of the field. Their results displayed that there was no significant difference in investing in defense more than offense and noted that a balance of both

was most likely to predict championships. While this article deals primarily with football, it shows the relevance of this question in sports as a whole and begs the question of whether their results hold true for a sport like basketball in which every player must play defense.

A more recent contribution to basketball literature is Otten and Miller (2015), which developed four hypotheses split between the NBA and NFL regarding offensive and defensive performance. For the NBA, the two hypotheses revolved around how individual and team statistics for defense correlate between the regular season and respective post season success. NBA statistics were gathered from 1950-2014 and teams were excluded if they did not qualify for the postseason. The main defensive statistics measured were opposing field goal percentage, opponent's average field goal percentage, opponent's points per game, and win percentage. Statistical techniques used in the analysis included ANOVA, MANOVA, multiple regression, and Pearson correlations. Results ultimately showed that the stated hypotheses received minor support and offensive and defensive field goal percentages were insignificant in predicting winning on their own, but were significant predictors when combined. This supports the Robst et al. (2011) article's conclusion that balanced teams are more successful than purely strong defensive teams. More importantly, these results are applicable on a professional level and directly involve past NBA seasons.

Despite the literature presented thus far, no study has closely examined the idea of defense winning championships in the modern NBA. The articles here have analyzed defensive basketball in many forms, from international to collegiate to professional, but today's NBA is more offensively focused than even that of 2015, the latest season studied in literature. An opening exists to determine which defensive statistics are reliable predictors of postseason success and if one of sports' oldest sayings holds true.

METHODS

To collect statistics from the past ten NBA seasons, data were drawn directly from Basketball Reference, a website that archives the officially recorded data from each NBA game and season and computes basic basketball analytical statistics using these data. Ten seasons worth of data were extracted for teams' regular season statistics, opponents' regular season statistics, and advanced statistics. Each of these statistical groupings were separate data sets within each year and were based on per game averages. The data were first sorted alphabetically by teams' names and then modified by adding variables indicating whether each team won that year's championship and whether each team reached its conference's championship series in the postseason. Variables, such as opponent field goals, opponent turnovers, opponent offensive rebounds, and defensive ratings, were then taken from the opponents' regular season statistics and advanced statistics and attached to the regular season statistics data sets, producing one data set for each year. These ten data sets were combined into one complete data set used to test the relationship between team defensive strength and winning championships. A data set containing leaguewide averages for the past decade of NBA seasons was also extracted to examine how 3-point field goals attempted and average points scored changed by season. R software, a free statistical programming language that is highly popular for data analysis, was used for the relevant computations. R software also has graphical capabilities that were utilized to display relationships and trends between variables.

The first statistical test performed compared the means of championship teams' defensive ratings to that of all non-championship teams. The resulting hypothesis test was

$$H_0: \mu_c - \mu_{nc} = 0 \quad vs \quad H_1: \mu_c - \mu_{nc} \neq 0$$

where μ_c is the mean defensive rating of championship teams over the past decade and μ_{nc} is the mean defensive rating of non-championship teams over the past decade. The designed hypotheses aimed directly on what impact a team's defensive performance had on championship success. A two-sample t-test was conducted at an $\alpha = 0.05$ significance level to test these hypotheses of whether the means of these two populations differed. For the results of the t-test to be reliable, assumptions of the continuity, normality, and variances of the population data sets had to be met. Other assumptions included that the samples are independent and simple random samples. The two-sample t-test was also used to construct a 95% confidence interval for the true difference in population means at an $\alpha = 0.05$ significance level.

Four multiple logistic regression models were then built to predict championship wins from a team's game statistics. Multiple logistic regression was designed to predict dichotomous outcomes, in this case whether a team won the championship or not, from multiple variables and provide probabilities of "success" for each observation. Hypothesis tests regarding logistic regression determine whether the predictor variables used actually influence the binary outcome. These hypothesis tests were conducted for each model and tested with the F-test of significance at an $\alpha = 0.05$ significance level. The Akaike information criterion (AIC) was used to compare the four models based on how well they fit the NBA data. The AIC gives a higher numeric score to models that do not predict the data as well as other possible models and penalizes models for each additional predictor variable. Each variable in each of the four models was tested for significance at an $\alpha = 0.05$ level to analyze which variables were the most significant predictors of winning an NBA championship. The assumptions of multiple logistic regression, a binary dependent variable, independence of observations, lack of correlation between independent

variables, linearity of independent variables, and large sample sizes, were checked for the champion and non-champion data sets.

The first multiple logistic regression model was based entirely on defensive statistics. This model aimed to determine how accurately a collection of an NBA team's defensive statistics could predict a championship success. The resulting model was

$$\ln\left[\frac{E(Y)}{1-E(Y)}\right] = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + x_8\beta_8 + x_9\beta_9$$

$$H_0: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_9 = 0 \quad vs \quad H_1: \beta_j \neq 0 \; for \; at \; least \; one \; j$$

where the x_j's were defensive statistics, the β_j's were unknown constants, and Y was a binary variable having values of 0 for a team that failed to win that year's championship, and 1 for a championship-winning team. The defensive statistics in this model were defensive rebounds per game (x_1), steals per game (x_2), blocks per game (x_3), personal fouls per game (x_4), opponent field goals per game (x_5), opponent free throws attempted per game (x_6), opponent offensive rebounds per game (x_7), opponent turnovers per game (x_8), and opponent three-point field goals per game (x_9).

The second multiple logistic regression model was also composed of solely defensive statistics, but only those deemed significant from previous research (Ibáñez et al., 2008). This model's goal was to ascertain whether only certain defensive statistics were needed to predict a championship team. The second model and set of hypothesis tests were

$$\ln\left[\frac{E(Y)}{1-E(Y)}\right] = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4$$

$$H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad vs \quad H_1: \beta_j \neq 0 \; for \; at \; least \; one \; j$$

with the same four defensive statistics as the first model: defensive rebounds, steals, blocks, and personal fouls, respectively.

An offensive statistics model was composed as the third multiple logistic regression model. The primary purpose of this model was to account for the impact that offense has on predicting a championship winning team. Statistics for this model were chosen based on their relevance to previously selected defensive statistics and their importance to the "modern NBA." This offensive model and hypothesis tests were

$$\ln\left[\frac{E(Y)}{1-E(Y)}\right] = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4$$

$$H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad vs \quad H_1: \beta_j \neq 0 \; for \; at \; least \; one \; j$$

with three-point field goals per game (x_1), offensive rebounds per game (x_2), free throws per game (x_3), and assists per game (x_4).

The fourth and final multiple logistic regression model was composed of a mix of offensive and defensive statistics. The statistics used for this model were combined from the statistics of the offensive and reduced defensive models. The final model and tests were

$$\ln\left[\frac{E(Y)}{1-E(Y)}\right] = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + x_8\beta_8$$

$$H_0: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_8 = 0 \quad vs \quad H_1: \beta_j \neq 0 \; for \; at \; least \; one \; j$$

where the x_j's are a combination of four offensive statistics and four defensive statistics from previous models.

To assess the accuracy of the models in terms of predicting NBA champions, leave-one-out cross-validation was conducted for each model. Cross-validation tests the precision of a model by splitting the relevant data into two sets: training data and test data. The chosen model has its rules for determining the Y variable outcome built on the training data and is then tested for accuracy with how well it predicts the test data. The purpose of this validation method is to test the model's performance on potential new data points. Leave-one-out cross-validation is a

variation of cross-validation where one observation is left out and treated as the test data set and the rest of the observations are used as the training data set. The advantage of the leave-one-out method is that models with relatively smaller data sets are still able to be reliably tested. The data set used for the four logistic regression models was 300 observations, but the champions and non-champions were unbalanced in there being 10 of the former and 290 of the latter. The performances of each model were then compared based on their accuracy scores.

The distinction between defense winning championships and defense putting teams in championship situations was determined through similar methods and models. A second hypothesis test regarding defensive ratings was conducted for teams that made their conference championship and for teams that failed to reach their conference championship:

$$H_0: \mu_{cc} - \mu_{ncc} = 0 \quad vs \quad H_1: \mu_{cc} - \mu_{ncc} \neq 0$$

with μ_{cc} as the mean defensive rating of conference championship teams and μ_{ncc} as the mean defensive rating of non-conference championship teams. A two-sample t-test was conducted at an $\alpha = 0.05$ significance level and the corresponding 95% confidence intervals were composed. The same four multiple logistic regression models and hypothesis tests for champions and non-champions were designed with the same setups and variables. The overall significance hypothesis tests were tested at an $\alpha = 0.05$ level and the models were compared based on their AIC scores. The offensive and reduced defensive variables from the second and third models were used to construct the fourth, mixed model. Each model was then further tested with leave-on-out cross-validation and the accuracy of each model was compared in terms of predicting new observations. These results were contrasted with the results for predicting championship teams with respect to how likely defensive statistics were to predict a team's postseason success in the modern NBA.

RESULTS AND DISCUSSION

The data for the defensive ratings of the champions and non-champions was found to be continuous as the ratings can take an uncountable range of values above zero. Further assumptions for normality of the two-sample t-test were tested with Quantile-Quantile, or Q-Q, plots (Appendix A). The championship-winning population ratings were found to be approximately normal, but this conclusion was limited by the small sample size (n = 10). Figure 2 shows the Q-Q plot for the defensive ratings of the champions.

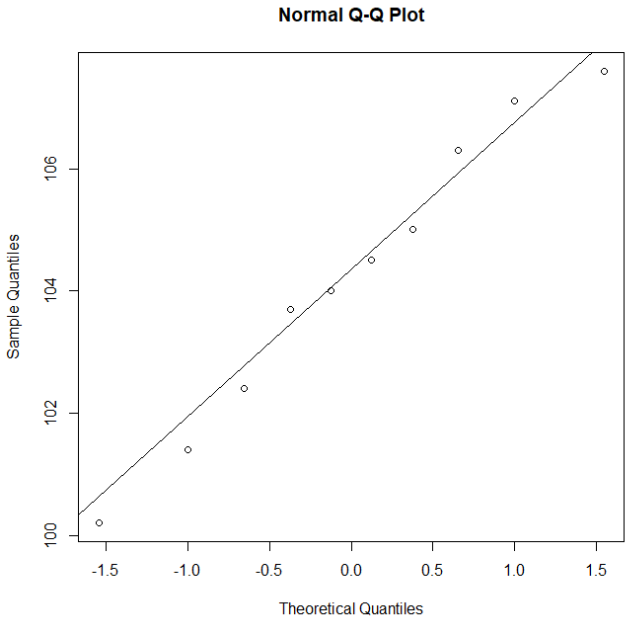


Figure 2. Q-Q plot of defensive ratings of championship teams

The non-champions' ratings were more clearly normal. The F-test about the equality of the population variances failed to reject the null hypothesis at $\alpha = 0.05$ with a p-value of 0.2238, and the variances could not be proved to be different. The two samples were not independent, as each of the championship teams faced teams from and had their defensive ratings influenced by the non-championship team sample. Teams playing in the same season were not independent of each other. Finally, the samples were not simple random samples because they were specifically

chosen based on a certain time frame, the past 10 years of the modern NBA. Based on the nature of this study, a simple random sample was impossible.

The two-sample t-test for the two populations returned a p-value of 0.0026, which rejected the null hypothesis at the $\alpha = 0.05$ significance level. The mean defensive ratings, therefore, were found to be significantly different with the champion population having the lower average mean. A lower defensive rating means a team allowed a fewer number of points per game than a higher rating, so the championship teams exhibited stronger defensive performances. The 95% Confidence Interval constructed, (-5.57, -1.19), supported this conclusion that the true difference between the population means was not 0. Championship teams in the modern NBA were estimated to have a mean defensive rating of 104.22 while non-championship teams were estimated to have a mean of 107.60. These championship teams played superior team defense and while this might not have been the only difference between the two populations, lower defensive ratings are a significant supporter of the idea that defense does win championships.

Assumptions of the multiple logistic regression models were checked for the champions and non-champions. The dependent variable, whether a team won that season's championship, was binary with outcomes of either 0 or 1. Independence of observations was violated for these models. Each teams' statistics depended on the other 29 teams for that season of play. Given the nature of basketball, several defensive and offensive statistics were highly correlated, which was another violation of these models. Figure 3 shows the correlations between statistics used across all four models. The size and color of circle between variables indicated how strongly the two variables were correlated.

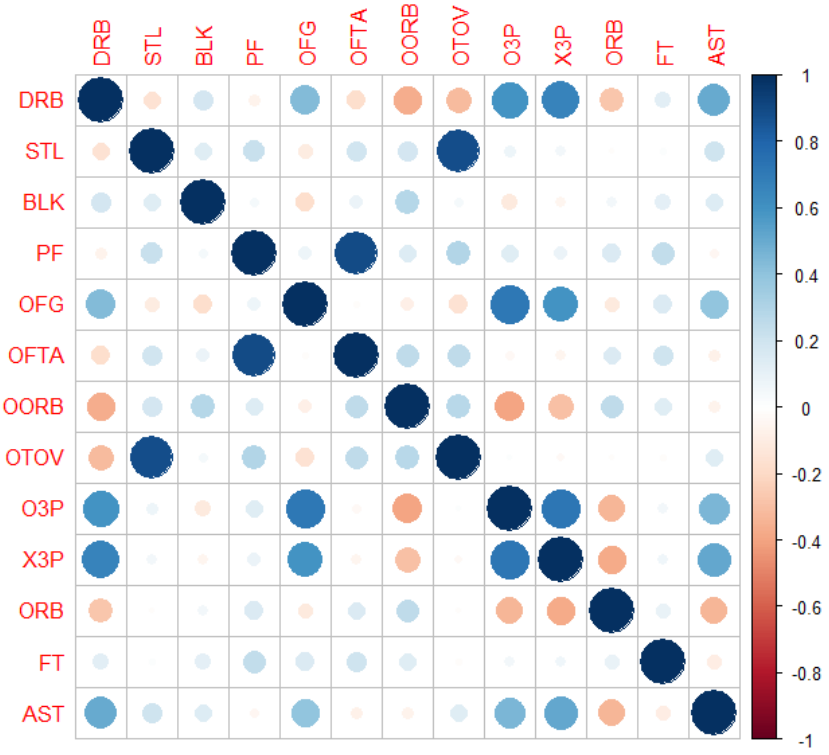


Figure 3. Correlations between offensive and defensive statistics

Steals (STL) and opponent turnovers (OTOV) were strongly positively correlated since a steal on the defensive side of the ball directly results in a turnover for the opposition. Personal fouls (PF) and opponent free throws attempted (OFTA) were highly correlated because a player usually is fouled before attempting a free throw shot. Opponent field goals (OFG) and defensive rebounds (DRB) were moderately correlated, as missing or scoring more field goals results in more or less rebounds for the opposition, respectively. The same relationship was found for defensive rebounds and opponent 3-point field goals (O3P). Opponent 3-point field goals and opponent field goals also had a strong correlation due to 3-point field goals being included in general field goals. Other significant correlations were found between 3-point field goals (X3P) and defensive rebounds, 3-point field goals and opponent 3-point field goals, 3-point field goals and assists (AST), opponent 3-point field goals and assists, and opponent offensive rebounds (OORB) and defensive rebounds. Each model, therefore, had correlated variables. The sample size of the non-

championship teams was large (n = 290), but the championship teams were still constrained by a small sample size.

The full defensive statistics model found none of the statistics used significant at an α=0.05 significance level. The AIC for this model was 87.009 and the null hypothesis of the β's was rejected at the $\alpha = 0.05$ level with a p-value of 0.014. This model suffered from overfitting and an excess of defensive statistics to the point that none were found to be significant predictors of a championship team. The correlation between variables likely had an impact on the significance of each variable. Despite this, the model was still found significant and supported the idea of a strong, all-around defense important in winning an NBA championship. Leave-one-out cross-validation with the full defensive statistics model produced an accuracy rate of 96.00%, meaning that 96 out of 100 potential future observations would be correctly classified according to their championship status. This model was highly accurate on defensive statistics alone, which supported the notion that defense is a reliable predictor of a championship team.

The second multiple logistic regression model with only four defensive statistics found both blocks and steals to be significant predictors of a championship team at $\alpha = 0.05$. The AIC for this reduced model was 80.910 and a p-value of 0.002 rejected the null hypothesis of the β's at this α. This model confirmed previous research that these four defensive statistics (defensive rebounds, blocks, steals, and personal fouls) were significant in championship success. Blocks and steals may have been significant predictors because steals can lead to fast break chances with high percentage field goals for teams and blocks directly prevent the opposition from scoring points. The cross-validation accuracy of this model was 96.67%. The reduced defensive statistics model focused on the most commonly thought of defensive statistics and showed a strong relationship between higher defensive statistics and championship teams.

At an $\alpha = 0.05$ significance level, the offensive statistics model found assists to be the sole significant predictor of winning a championship in the NBA. This model's AIC was 80.076. The null hypothesis of the β's for this model at $\alpha = 0.05$ was rejected with a p-value of 0.001; $\beta_j \neq 0$ for at least one j. Unsurprisingly, offensive statistics were significant in regards to championship success in an offensively-dominated league. Assists were likely the only significant predictor because they are guaranteed to precede a made field goal, which could either be for 2 or 3 points. The offensive model correctly predicted 96.67% of observations with the cross-validation implemented. Championship teams rely on their great offenses to score enough points to win games.

In the fourth multiple logistic regression model, built with the offensive and defensive statistics of models two and three, no variables were found to be significant at the $\alpha = 0.05$ significance level. A lack of significant statistics again could be related to the correlation between the variables in this model. The model had an AIC of 82.357 and a p-value of 0.003 for overall significance. The null hypothesis of the β's was rejected, and the variables within the model were found significant for predicting a championship result. Balanced teams, teams that had strong offensive and defensive performances, have had notable postseason success over the past decade. The mixed offensive and defensive model performed well with cross-validation at a 97.33% accuracy rate.

All four models were significant in predicting championship teams, but the models did not perform equally well. The offensive model fit the data the best with the lowest AIC, but the reduced defensive model had a close second best fit. The first model had the highest AIC and was likely overloaded with correlated variables, a problem that the reduced model fixed. The mixed model had the best accuracy rate for predicting whether a team won that season's

championship, and the accuracy rates for the offensive and reduced defensive models were the same. Superior offensive and defensive performances on their own were indicative of championship success, but the combination of the two was better suited for predicting whether a new observation won a championship. The added context of offense as well as defense in a single model made a difference in estimated accuracies and likely provided more information for the models to be trained on. Like the sport of basketball itself, offense and defense are both needed to build the most accurate model of what a championship team looks like. Still, defense plays an important role in winning a championship in the modern NBA.

The conference championship teams were then tested to compare results. The defensive ratings for conference championship teams were still continuous. Non-conference championship teams' ratings were found to be normal, but those of conference championship teams were limited in their appearance of normalcy (Appendix B). Figure 4 shows the Q-Q plot of the conference championship teams.

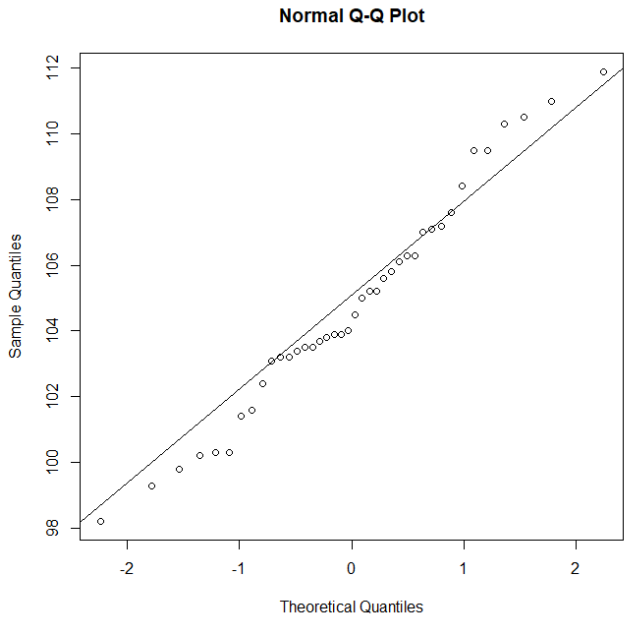


Figure 4. Q-Q plot of defensive ratings of conference championship teams

The data was limited by the relatively small sample size (n = 40) but large enough to discern that the points do not exactly follow the normal shape, but were close to normal. The population variances of the conference championship and non-conference championship teams could not be proved unequal from the F-test at an $\alpha = 0.05$ level. The two samples were not independent or simple random samples due to the nature of this study.

The two-sample t-test for conference championship participants and non-participants returned a p-value close to 0 and rejected the null hypothesis at $\alpha = 0.05$. Teams that made their conference's championship series had significantly higher defensive ratings than those that did not. The true difference between the mean defensive ratings of these populations is, with 95% confidence, between -4.196 and -1.964. This interval did not include 0 and supported defense putting teams into a championship situation. Defensive ratings are not indicative of a team's overall performance, but these results showed the significance of defensive performance in making a deep postseason run. NBA teams over the past decade have utilized their defensive skills to put themselves in positions to win some type of championship.

The assumptions of the multiple logistic regression models for conference championship participants and non-participants resembled those of the championship and non-championship teams. The dependent variable, a team's conference championship status, had dichotomous outcomes of 0 and 1. Independence of observations was not achieved. Defensive and offensive statistics alike were highly correlated across all four models. Non-participants of the conference championship had a large sample size (n = 260) but there were only 40 participants, which caused an imbalance between the successes and failures. Therefore, not all of the assumptions of multiple logistic regression could be met.

The first model with full defensive statistics found steals and opponent field goals significant predictors of conference championship participation at $\alpha = 0.05$. The significance of opponent field goals suggested that defense that is not directly recorded in statistics, such as contesting opponents' shots, is important in reaching the conference championship. This model's AIC was 207.54. The overall test of significance rejected the null hypothesis of the β's with a p-value close to 0. Defense, as it was for winning the NBA championship, was important to teams extending their playoff success. Leave-one-out cross-validation with this model produced an accuracy rate of 85.33%. This model struggled more with predicting conference championship teams than it did championship teams but still maintained a decently high rate.

The reduced defensive statistics model found all four statistics significant at $\alpha = 0.05$. This suggests that strong defense, especially through these main statistics, is crucial to a team's chances of reaching the conference championship. More defensive rebounds, blocks, and steals and less personal fouls were found in these top four postseason teams. The reduced model's AIC was 215.30 and the overall test of significance rejected the null hypothesis of the β's with a p-value of practically 0. Cross-validation showed that the reduced model was 86.67% successful in predicting future observations. This model also supported the idea that defense plays an important role in reaching a championship situation.

The third, offensive model tested at $\alpha = 0.05$ found offensive rebounds, free throws, and assists to be significant predictors of conference championship participation. The only variable not found to be significant was 3-point field goals, despite the recent offensive trend in the NBA. Due to the proved correlation between 3-point field goals and assists, it is possible this prevented 3-point shooting from being a significant predictor. Offensive rebounds produce more field goal chances for teams, free throws typically get the opposition into foul trouble as well as providing

uncontested shots, and assists increase a team's ball movement, leading to better shots. All of these combined allow a team to score more points and win more games. The offensive model's AIC was 223.29 and the null hypothesis of the β's was once again rejected with a p-value of almost 0. The model performed well under cross-validation with an accuracy rate of 86.67%. Offensive performance proved to be useful in assessing whether a team reached its conference championship.

The final model found blocks, personal fouls, and free throws as significant predictors at an $\alpha = 0.05$ level. A mix of one offensive statistic and two defensive statistics being indicative of conference championship participation showed the importance of a team being solid on both sides of the court. The mixed model had an AIC of 210.76. The overall test of significance produced a p-value of practically 0, therefore rejecting the null hypothesis of the β's at the given significance level. An accuracy rate of 86.33% from the mixed model was obtained with the leave-one-out cross-validation. As was the case for championship teams, a balance of offense and defense was significant in both predicting potential future conference championship participants and in the teams that made their conference's championship in the past decade of play.

Each model had its own strengths. The full defensive statistics model fit the data the best given it had the lowest AIC. The reduced defensive and solely offensive models produced the highest accuracy rates of predicting future conference championship participants. The mixed model had decent overall performance as it fit the data relatively well and maintained a high accuracy rate. Many variables of both offensive and defensive nature proved to be significant within the contexts of the different models. Defensive performance is undoubtedly an important factor in a team reaching the conference championship, but these results are unable to distinguish

whether defense alone is more important for a deep postseason run than offense or both offense and defense combined.

The four multiple logistic regression models performed differently with regards to championship teams and conference championship teams. The accuracy rates for participation in conference championships were lower, but more individual statistics were significant class predictors. Both series of tests, however, were consistent in defense and offense being level in terms of predicting championship and championship situation success. The separate models were successful on their own, but neglected that, on the court, neither offense nor defense can survive on its own. In both cases, the mixed models were able to perform about as equally well as the individual models in predictions and fitting of the data. The conference championship and NBA championship teams may have been stronger on one side of the court than the other, but they were still usually superior in both aspects than the other teams. The defense rating tests further proved that defense has still been a significant feature of the most successful teams in what has been coined the modern NBA.

Several limitations were imposed upon the data, models, and testing. One of the most notable limitations was the lack of data for championship teams. Since increased 3-point field goals attempted and overall scoring is a recent trend in the NBA's history, only 10 championship teams were observed. The created a large imbalance between the success and failures of the logistic regression. Two of the observed seasons also experienced fewer total games than the usual 1,230 games played across a season. The 2011-2012 regular season experienced a lockout and the 2019-2020 regular season experienced the global COVID-19 pandemic, both of which shortened the amount of games played. The regular season statistics were used due to being the largest sample size of games for all teams, but regular season success does not guarantee the

same level of play within the playoffs. Oftentimes, lower seeded teams heighten their level of play in the postseason and make unexpected runs. Further limitations were found in the models and tests used for the data. Two-sample t-tests and logistic regression require independence of observations in their assumptions, but the majority of basketball data cannot be independent due to a team's statistics depending on their play against other teams. A simple random sample may have been possible for this 10 year timeframe, but this would have caused further problems in the already limited sample size. Leave-one-out cross-validation faced limitations within this study, as well. The imbalance of the championship and non-championship teams meant that if the model predicted every "new" observation as a non-championship team, the model would still maintain an accuracy rate of 96.67%, which may have happened for one of the models in this study.

CONCLUSION

Defense, while still an important contributor, is not the sole reason that teams in the past 10 years of an offensively-oriented NBA have won championships. Defensive ratings of championship teams and conference championship participants are significantly lower than average, but multiple logistic regression models composed of offensive and defensive statistics did not largely differ in their fit to the data and their accuracy in predicting future observations. Models with both offensive and defensive statistics combined displayed a more cohesive idea of superiority in both categories being an integral part in championship performances. Much like in today's NFL, the NBA has seen its most successful teams dominate on both sides of the ball (Robst, VanGlider, Berri, and Vance, 2011). NBA champions of the coming years are likely to be best predicted by their combined offensive and defensive performances rather than on defense alone.

The phrase "defense wins championships" is not entirely true in the modern NBA as attributing championship solely to defense is misleading. Defense has gotten teams to championships and championship situations, but focusing on this aspect undersells the equally important impact offense has had in earning teams titles. Therefore, this sports axiom should be revisited and revised in the context of today's NBA. Perhaps "defense wins championships in collaboration with strong offense" or "defense wins championships but not without offensive assistance" are more honest depictions of the most recent championship teams. Basketball has evolved offensively to the point where great defense cannot carry a team to a trophy on its own. Defense still holds an important role in the modern NBA, but it should not be credited for the entirety of postseason success.

Further research should look for similar modern offensive trends in women's, college, and international basketball leagues and analyze the impact defense has on championships in these different contexts. More logistic regression models with both offensive and defensive statistics should be compared with exclusively defensive models to determine which are better predictors of championship-winning teams. To compare prediction accuracy rates, other cross-validation methods should be experimented with to avoid inflated rates due to overestimation of non-champions. Further studies of defense and championships in the modern NBA should involve the collection of more data as future seasons are completed and offensive strategies continue to develop. Defensive performances of championship teams should also be contrasted with those of playoff teams to examine the difference defense creates within varying levels of postseason success.

REFERENCES

Álvarez, A., Ortega, E., Gómez, M. A., & Salado, J. (2009). Study of the Defensive Performance Indicators in Peak Performance Basketball. *Revista de Psicología del Deporte*, 18, 379-384. Retrieved November 6, 2020, from https://revistes.uab.cat/rpd/article/view/642/608

Cappe. (2020, September 3). *Learn a stat: Strength of schedule (sos)*. https://hackastat.eu/en/learn-a-stat-strength-of-schedule-sos/.

Foxworth, D. (2018, December 8). *It's time we modify old adage 'defense Wins championships'*. https://theundefeated.com/features/its-time-we-modify-old-adage-defense-wins-championships/.

Ibáñez, S. J., Sampaio, J., Feu, S., Lorenzo, A., Gómez, M. A., & Ortega, E. (2008). Basketball game-related statistics that discriminate between teams' season-long success. *European Journal of Sport Science, 8*(6), 369-372. https://doi.org/10.1080/17461390802261470

Kotecki, J. (2014) Estimating the Effect of Home Court Advantage on Wins in the NBA. *The Park Place Economist, 22*(1), 49-57.

Mondello, M. (2000). Does Defense Win Championships? *Strategies, 13*(3), 34-36. https://doi.org/10.1080/08924562.2000.10591440

NBA League Averages – Per Game. (n.d.). Basketball Reference. Retrieved November 6, 2020, from https://www.basketball-reference.com/leagues/NBA_stats_per_game.html

Nweiyue. (2020, August 6). *NBA: Offense win games, defense win championships. Is this really the case?* Medium. https://medium.com/@nweiyue/nba-offense-win-games-defense-win-championships-is-this-really-the-case-6faefe6d9c9b.

Otten, M. P., & Miller, T. J. (2015). A Balanced Team Wins Championships: 66 Years of Data from the National Basketball Association and the National Football League. *Perceptual*

& Motor Skills: Exercise & Sport, 121(3), 654-665. https://doi.org/10.2466/30.26.PMS.121c25x4

Robst, J., VanGlider, J., Berri, D. J., & Vance, C. (2011). 'Defense Wins Championships?' The Answer from the Gridiron. *International Journal of Sport Finance, 6*(1), 72-84. Retrieved November 6, 2020, from https://search.proquest.com/openview/d8dd45539ffbe156ca7df12f9e2bedad/1?pq-origsite=gscholar&cbl=28340

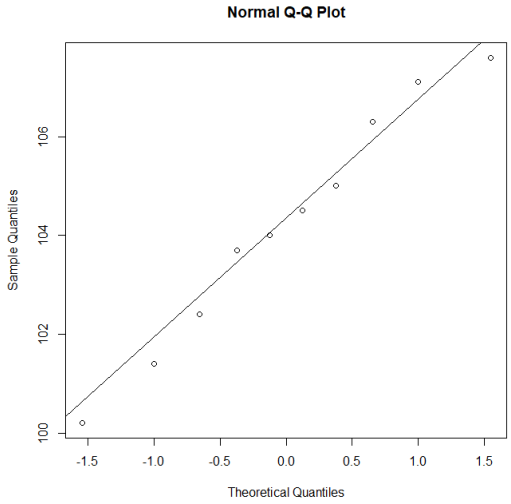
Shea, S. (2018). *The 3-Point Revolution*. Shottracker. https://shottracker.com/articles/the-3-point-revolution.

USA Men's National Teams. U.S. Olympic Men's Basketball Teams. (2021). https://www.usab.com/mens/national-team.aspx.

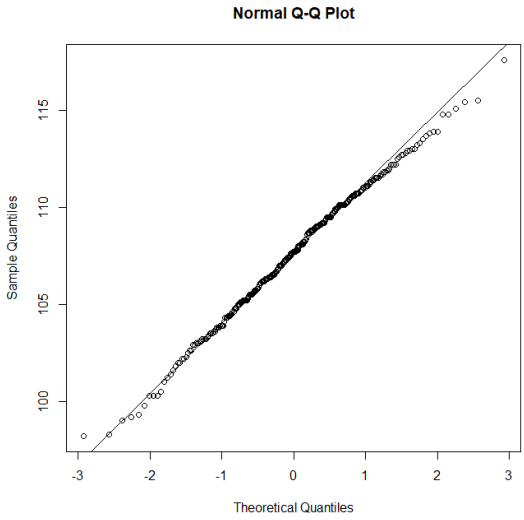
Zuccolotto, P., & Manisera, M. (2020). *Basketball data science with applications in r.* Taylor & Francis Group.

APPENDIX A

Assumption of Normality Test for Defensive Ratings of Championship and Non-Championship Populations



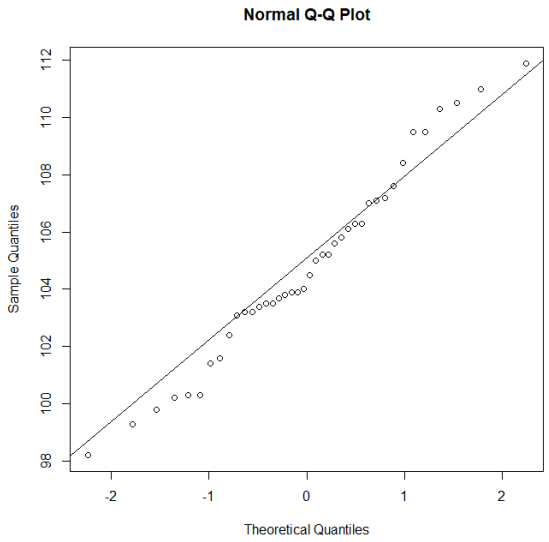
Q-Q plot of championship population



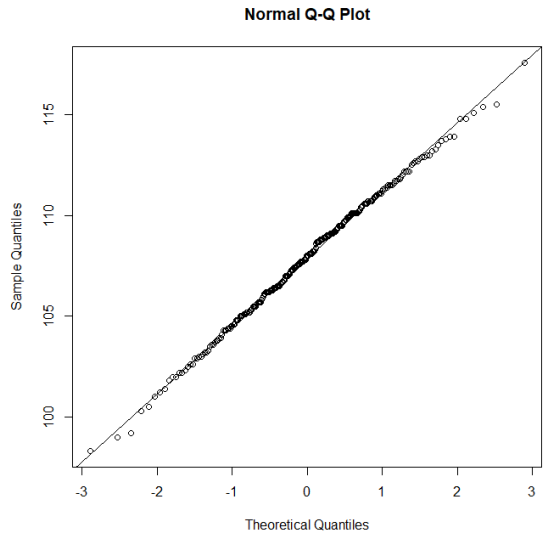
Q-Q plot of non-championship population

APPENDIX B

Assumption of Normality Test for Defensive Ratings of Conference Championship Participants and Non-Participant Populations



Q-Q plot of conference championship participants



Q-Q plot of conference championship non-participants

APPENDIX C

R Software Code for Models and Methods

```
# Reading in the data sets
# Change directory first
leaguedata <- read.csv(file='leaguedata.csv', header=T)
# Sub setting this data set into the “modern NBA”
leaguedata <- leaguedata[2:11,]
attach(leaguedata)
leaguedata$Season <- c(2019,2018,2017,2016,2015,2014,2013,2012,2011,2010)
leaguedata <- leaguedata[order(leaguedata$Season),]

# Reading in 2010-2011 data sets
detach(leaguedata)
data2010 <- read.csv(file='2010-11data.csv', header=T)
champion <- numeric(31)
champion <- as.numeric(data2010$Team == "Dallas Mavericks*")
top4 <- numeric(31)
top4 <- as.numeric(data2010$Team == "Dallas Mavericks*" | data2010$Team == "Miami
Heat*" | data2010$Team == "Chicago Bulls*" | data2010$Team == "Oklahoma City Thunder*")
data2010 <- cbind(data2010, champion, top4)
data2010 <- data2010[-31,]
data2010 <- data2010[order(data2010$Team),]
oppdata2010 <- read.csv(file='2010-11oppdata.csv', header=T)
oppdata2010 <- oppdata2010[-31,]
oppdata2010 <- oppdata2010[order(oppdata2010$Team),]
attach(oppdata2010)
data2010 <- cbind(data2010, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2010)
miscdata2010 <- read.csv(file='2010-11miscdata.csv', header=T)
miscdata2010 <- miscdata2010[-c(31,32),]
miscdata2010 <- miscdata2010[order(miscdata2010$Team),]
attach(miscdata2010)
data2010 <- cbind(data2010, DRtg)
detach(miscdata2010)

#Reading in 2011-2012 data sets
data2011 <- read.csv(file='2011-12data.csv', header=T)
champion <- numeric(31)
champion <- as.numeric(data2011$Team == "Miami Heat*")
top4 <- numeric(31)
top4 <- as.numeric(data2011$Team == "Boston Celtics*" | data2011$Team == "Miami Heat*" |
data2011$Team == "San Antonio Spurs*" | data2011$Team == "Oklahoma City Thunder*")
data2011 <- cbind(data2011, champion, top4)
data2011 <- data2011[-31,]
```

```
data2011 <- data2011[order(data2011$Team),]
oppdata2011 <- read.csv(file='2011-12oppdata.csv', header=T)
oppdata2011 <- oppdata2011[-31,]
oppdata2011 <- oppdata2011[order(oppdata2011$Team),]
attach(oppdata2011)
data2011 <- cbind(data2011, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2011)
miscdata2011 <- read.csv(file='2011-12miscdata.csv', header=T)
miscdata2011 <- miscdata2011[-c(31,32),]
miscdata2011 <- miscdata2011[order(miscdata2011$Team),]
attach(miscdata2011)
data2011 <- cbind(data2011, DRtg)
detach(miscdata2011)

#Reading in 2012-2013 data sets
data2012 <- read.csv(file='2012-13data.csv', header=T)
champion <- numeric(31)
champion <- as.numeric(data2012$Team == "Miami Heat*")
top4 <- numeric(31)
top4 <- as.numeric(data2012$Team == "Indiana Pacers*" | data2012$Team == "Miami Heat*" |
data2012$Team == "San Antonio Spurs*" | data2012$Team == "Memphis Grizzlies*")
data2012 <- cbind(data2012, champion, top4)
data2012 <- data2012[-31,]
data2012 <- data2012[order(data2012$Team),]
oppdata2012 <- read.csv(file='2012-13oppdata.csv', header=T)
oppdata2012 <- oppdata2012[-31,]
oppdata2012 <- oppdata2012[order(oppdata2012$Team),]
attach(oppdata2012)
data2012 <- cbind(data2012, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2012)
miscdata2012 <- read.csv(file='2012-13miscdata.csv', header=T)
miscdata2012 <- miscdata2012[-c(31,32),]
miscdata2012 <- miscdata2012[order(miscdata2012$Team),]
attach(miscdata2012)
data2012 <- cbind(data2012, DRtg)
detach(miscdata2012)

#Reading in 2013-2014 data sets
data2013 <- read.csv(file='2013-14data.csv', header=T)
champion <- numeric(31)
champion <- as.numeric(data2013$Team == "San Antonio Spurs*")
top4 <- numeric(31)
top4 <- as.numeric(data2013$Team == "Indiana Pacers*" | data2013$Team == "Miami Heat*" |
data2013$Team == "San Antonio Spurs*" | data2013$Team == "Oklahoma City Thunder*")
data2013 <- cbind(data2013, champion, top4)
data2013 <- data2013[-31,]
```

```
data2013 <- data2013[order(data2013$Team),]
oppdata2013 <- read.csv(file='2013-14oppdata.csv', header=T)
oppdata2013 <- oppdata2013[-31,]
oppdata2013 <- oppdata2013[order(oppdata2013$Team),]
attach(oppdata2013)
data2013 <- cbind(data2013, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2013)
miscdata2013 <- read.csv(file='2013-14miscdata.csv', header=T)
miscdata2013 <- miscdata2013[-c(31,32),]
miscdata2013 <- miscdata2013[order(miscdata2013$Team),]
attach(miscdata2013)
data2013 <- cbind(data2013, DRtg)
detach(miscdata2013)

#Reading in 2014-2015 data sets
data2014 <- read.csv(file='2014-15data.csv', header=T)
champion <- numeric(31)
champion <- as.numeric(data2014$Team == "Golden State Warriors*")
top4 <- numeric(31)
top4 <- as.numeric(data2014$Team == "Golden State Warriors*" | data2014$Team ==
"Cleveland Cavaliers*" | data2014$Team == "Atlanta Hawks*" | data2014$Team == "Houston
Rockets*")
data2014 <- cbind(data2014, champion, top4)
data2014 <- data2014[-31,]
data2014 <- data2014[order(data2014$Team),]
oppdata2014 <- read.csv(file='2014-15oppdata.csv', header=T)
oppdata2014 <- oppdata2014[-31,]
oppdata2014 <- oppdata2014[order(oppdata2014$Team),]
attach(oppdata2014)
data2014 <- cbind(data2014, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2014)
miscdata2014 <- read.csv(file='2014-15miscdata.csv', header=T)
miscdata2014 <- miscdata2014[-c(31,32),]
miscdata2014 <- miscdata2014[order(miscdata2014$Team),]
attach(miscdata2014)
data2014 <- cbind(data2014, DRtg)
detach(miscdata2014)

#Reading in 2015-2016 data sets
data2015 <- read.csv(file='2015-16data.csv', header=T)
champion <- numeric(31)
champion <- as.numeric(data2015$Team == "Cleveland Cavaliers*")
top4 <- numeric(31)
top4 <- as.numeric(data2015$Team == "Cleveland Cavaliers*" | data2015$Team == "Golden
State Warriors*" | data2015$Team == "Toronto Raptors*" | data2015$Team == "Oklahoma City
Thunder*")
```

```
data2015 <- cbind(data2015, champion, top4)
data2015 <- data2015[-31,]
data2015 <- data2015[order(data2015$Team),]
oppdata2015 <- read.csv(file='2015-16oppdata.csv', header=T)
oppdata2015 <- oppdata2015[-31,]
oppdata2015 <- oppdata2015[order(oppdata2015$Team),]
attach(oppdata2015)
data2015 <- cbind(data2015, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2015)
miscdata2015 <- read.csv(file='2015-16miscdata.csv', header=T)
miscdata2015 <- miscdata2015[-c(31,32),]
miscdata2015 <- miscdata2015[order(miscdata2015$Team),]
attach(miscdata2015)
data2015 <- cbind(data2015, DRtg)
detach(miscdata2015)

#Reading in 2016-2017 data sets
data2016 <- read.csv(file='2016-17data.csv', header=T)
champion <- numeric(31)
champion <- as.numeric(data2016$Team == "Golden State Warriors*")
top4 <- numeric(31)
top4 <- as.numeric(data2016$Team == "Golden State Warriors*" | data2016$Team ==
"Cleveland Cavaliers*" | data2016$Team == "Boston Celtics*" | data2016$Team == "San
Antonio Spurs*")
data2016 <- cbind(data2016, champion, top4)
data2016 <- data2016[-31,]
data2016 <- data2016[order(data2016$Team),]
oppdata2016 <- read.csv(file='2016-17oppdata.csv', header=T)
oppdata2016 <- oppdata2016[-31,]
oppdata2016 <- oppdata2016[order(oppdata2016$Team),]
attach(oppdata2016)
data2016 <- cbind(data2016, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2016)
miscdata2016 <- read.csv(file='2016-17miscdata.csv', header=T)
miscdata2016 <- miscdata2016[-c(31,32),]
miscdata2016 <- miscdata2016[order(miscdata2016$Team),]
attach(miscdata2016)
data2016 <- cbind(data2016, DRtg)
detach(miscdata2016)

# Reading in 2017-2018 data sets
data2017 <- read.csv(file='2017-18data.csv', header=T)
champion <- numeric(31)
champion <- as.numeric(data2017$Team == "Golden State Warriors*")
top4 <- numeric(31)
```

```
top4 <- as.numeric(data2017$Team == "Golden State Warriors*" | data2017$Team ==
"Cleveland Cavaliers*" | data2017$Team == "Boston Celtics*" | data2017$Team == "Houston
Rockets*")
data2017 <- cbind(data2017, champion, top4)
data2017 <- data2017[-31,]
data2017 <- data2017[order(data2017$Team),]
oppdata2017 <- read.csv(file='2017-18oppdata.csv', header=T)
oppdata2017 <- oppdata2017[-31,]
oppdata2017 <- oppdata2017[order(oppdata2017$Team),]
attach(oppdata2017)
data2017 <- cbind(data2017, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2017)
miscdata2017 <- read.csv(file='2017-18miscdata.csv', header=T)
miscdata2017 <- miscdata2017[-c(31,32),]
miscdata2017 <- miscdata2017[order(miscdata2017$Team),]
attach(miscdata2017)
data2017 <- cbind(data2017, DRtg)
detach(miscdata2017)

# Reading in 2018-2019 data sets
data2018 <- read.csv(file='2018-19data.csv', header=T)
champion <- numeric(31)
champion <- as.numeric(data2018$Team == "Toronto Raptors*")
top4 <- numeric(31)
top4 <- as.numeric(data2018$Team == "Golden State Warriors*" | data2018$Team == "Toronto
Raptors*" | data2018$Team == "Milwaukee Bucks*" | data2018$Team == "Portland Trail
Blazers*")
data2018 <- cbind(data2018, champion, top4)
data2018 <- data2018[-31,]
data2018 <- data2018[order(data2018$Team),]
oppdata2018 <- read.csv(file='2018-19oppdata.csv', header=T)
oppdata2018 <- oppdata2018[-31,]
oppdata2018 <- oppdata2018[order(oppdata2018$Team),]
attach(oppdata2018)
data2018 <- cbind(data2018, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2018)
miscdata2018 <- read.csv(file='2018-19miscdata.csv', header=T)
miscdata2018 <- miscdata2018[-c(31,32),]
miscdata2018 <- miscdata2018[order(miscdata2018$Team),]
attach(miscdata2018)
data2018 <- cbind(data2018, DRtg)
detach(miscdata2018)

# Reading in 2019-2020 data sets
data2019 <- read.csv(file='2019-20data.csv', header=T)
champion <- numeric(31)
```

```
champion <- as.numeric(data2019$Team == "Los Angeles Lakers*")
top4 <- numeric(31)
top4 <- as.numeric(data2019$Team == "Los Angeles Lakers*" | data2019$Team == "Miami
Heat*" | data2019$Team == "Boston Celtics*" | data2019$Team == "Denver Nuggets*")
data2019 <- cbind(data2019, champion, top4)
data2019 <- data2019[-31,]
data2019 <- data2019[order(data2019$Team),]
oppdata2019 <- read.csv(file='2019-20oppdata.csv', header=T)
oppdata2019 <- oppdata2019[-31,]
oppdata2019 <- oppdata2019[order(oppdata2019$Team),]
attach(oppdata2019)
data2019 <- cbind(data2019, OFG, OFTA, OORB, OTOV, O3P, O3PA)
detach(oppdata2019)
miscdata2019 <- read.csv(file='2019-20miscdata.csv', header=T)
miscdata2019 <- miscdata2019[-c(31,32),]
miscdata2019 <- miscdata2019[order(miscdata2019$Team),]
attach(miscdata2019)
data2019 <- cbind(data2019, DRtg)
detach(miscdata2019)

# Putting the yearly data sets together into one dataframe
rm(champion)
rm(top4)
alldata <- rbind(data2010, data2011, data2012, data2013, data2014, data2015, data2016,
data2017, data2018, data2019)

# Scatterplot for Season, PTS, 3PA
symbols(leaguedata$Season, leaguedata$PTS, circles=leaguedata$X3PA,inches=0.1,
xlab="Season", ylab="Average Points Scored", main="Average Points Scored per NBA Season
Accounting for 3PA")

# Hypothesis testing for defensive ratings of champions
# Assumptions
attach(alldata)
champions <- alldata[champion == 1,]
nonchampions <- alldata[champion == 0,]
qqnorm(champions$DRtg)
qqline(champions$DRtg)
qqnorm(nonchampions$DRtg)
qqline(nonchampions$DRtg)
var.test(champions$DRtg, nonchampions$DRtg)
# Actual t-test
t.test(champions$DRtg, nonchampions$DRtg, var.equal=TRUE)

# Correlations between variables in suggested models
library(corrplot)
```

```
vars.model <- cbind(DRB, STL, BLK, PF, OFG, OFTA, OORB, OTOV, O3P, X3P, ORB, FT,
AST)
corrplot(cor(vars.model), method="circle")

#Multiple logistic regression
model.full <- glm(champion ~ DRB + STL + BLK + PF + OFG + OFTA + OORB +
OTOV+O3P, family = binomial)
model.prev4 <- glm(champion ~ DRB + STL + BLK + PF , family = binomial)
model.off <- glm(champion ~ X3P + ORB + FT + AST, family = binomial)
model.mixed <- glm(champion ~ DRB + STL + BLK + PF + X3P + ORB + FT + AST , family =
binomial)
summary(model.full)
1-pchisq(87.687-67.009, 299-290)
summary(model.prev4)
1-pchisq(87.687-70.910, 299-295)
summary(model.off)
1-pchisq(87.687-70.076, 299-295)
summary(model.mixed)
1-pchisq(87.687- 64.357, 299-291)

# Cross-validation
library(caret)
train.control <- trainControl(method = "LOOCV")
champ.factor <- as.factor(champion)
alldata <- cbind(alldata, champ.factor)
cvmodel.full <- train(champ.factor ~ DRB + STL + BLK + PF + OFG + OFTA + OORB +
OTOV + O3P, data = alldata, method = "glm", trControl = train.control)
cvmodel.full
head(cvmodel.full)
cvmodel.prev4 <- train(champ.factor ~ DRB + STL + BLK + PF, data = alldata, method =
"glm", trControl = train.control)
cvmodel.prev4
head(cvmodel.prev4)
cvmodel.off <- train(champ.factor ~ X3P + ORB + FT + AST, data = alldata, method = "glm",
trControl = train.control)
cvmodel.off
cvmodel.mixed <- train(champ.factor ~ DRB + STL + BLK + PF + X3P + ORB + FT + AST,
data = alldata, method = "glm", trControl = train.control)
cvmodel.mixed

#Hypothesis Test for Conference Championship Participants
confchamp <- alldata[top4 == 1,]
nonconfchamp <- alldata[top4 == 0,]
qqnorm(confchamp$DRtg)
qqline(confchamp$DRtg)
qqnorm(nonconfchamp$DRtg)
```

```
qqline(nonconfchamp$DRtg)
var.test(confchamp$DRtg, nonconfchamp$DRtg)
# Actual t-test
t.test(confchamp$DRtg, nonconfchamp$DRtg, var.equal=TRUE)

# Multiple Logistic Regression
model.full.cc <- glm(top4 ~ DRB + STL + BLK + PF + OFG + OFTA + OORB + OTOV+O3P,
family = binomial)
model.prev4.cc <- glm(top4 ~ DRB + STL + BLK + PF , family = binomial)
model.off.cc <- glm(top4 ~ X3P + ORB + FT + AST, family = binomial)
model.mixed.cc <- glm(top4 ~ DRB + STL + BLK + PF + X3P + ORB + FT + AST, family =
binomial)
summary(model.full.cc)
1-pchisq(235.60-187.54,299-290)
summary(model.prev4.cc)
1-pchisq(235.6-205.3, 299-295)
summary(model.off.cc)
1-pchisq(235.60-213.29, 299-295)
summary(model.mixed.cc)
1-pchisq(235.60-192.76, 299-291)

# Cross-validation
top4.factor <- as.factor(top4)
alldata <- cbind(alldata, top4.factor)
cvmodel.full.cc <- train(top4.factor ~ DRB + STL + BLK + PF + OFG + OFTA + OORB +
OTOV + O3P, data = alldata, method = "glm", trControl = train.control)
cvmodel.full.cc
cvmodel.prev4.cc <- train(top4.factor ~ DRB + STL + BLK + PF, data = alldata, method =
"glm", trControl = train.control)
cvmodel.prev4.cc
cvmodel.off.cc <- train(top4.factor ~ X3P + ORB + FT + AST, data = alldata, method = "glm",
trControl = train.control)
cvmodel.off.cc
cvmodel.mixed.cc <- train(top4.factor ~ DRB + STL + BLK + PF + X3P + ORB + FT + AST,
data = alldata, method = "glm", trControl = train.control)
cvmodel.mixed.cc
```