5-17-2022

An Econometric Analysis of Collegiate Player Performance to Create a Model for Forecasting Contributions to Team Success

Evan Seely
evan.seely@comcast.net

An Econometric Analysis of Collegiate Player Performance to Create a Model for Forecasting Contributions to Team Success

Evan Seely

Bellarmine University

Honors Program

Dr. Michael Ackerman, Advisor

Dr. Frank Raymond, Reader

April 2022

Table of Contents

List of Tables

Abstract

At the conclusion of each basketball season, each conference selects 1st, 2nd, and sometimes 3rd all-conference teams based on player performance for that season. Often, these all-conference teams reflect biases in the media rather than evaluations based on player performance alone. The baseball statistic Wins Above Replacement, WAR, is useful in quantifying the impact of each player through the number of wins contributed to his respective team by comparing each player to a designated replacement level player. This statistic can also be applied to basketball analysis to perform a similar function as in baseball, despite a vastly different formulation. However, the WAR statistic has limitations in its player analysis in basketball, particularly through failing to include defensive statistics and having no established definition of a replacement player. In this paper, I utilize the Wins Above Replacement statistic along with other key statistics, particularly in the defensive aspect of the game, to create an econometric model to better determine which players contributed the most to their team's success. These statistics determine which players should be selected to the all-conference team at the end of the collegiate basketball season.

Introduction

At the conclusion of each basketball season, each conference selects the all-conference teams that consist of the top players based on the performances of the recently completed season. The smaller conferences typically choose a first and second team (ASUN Conference, 2020) with the larger conferences selecting a first, second, and third team (ACC Network, 2021). Additionally, all conferences typically select an all-freshman team and an all-defensive team along with honoring a conference player of the year, a defensive player of the year, and a coach of the year (ASUN Conference, 2020). These teams and accolades are chosen by a combination of head coaches in the conference and members of the media who covered the conference throughout the year. Too often, these selections do not result in the truly deserving players, who had the best statistical seasons, being selected to the all-conference team. These negligent selections can be attributed to several different biases such as narratives created by the media, recency bias, which members of the media are allowed to vote, bias toward winning teams, or other discrepancies that arise from the voting process. Additionally, the criterion for selection is not always clear whether the committee's methods are based upon on court performance or a perceived image off the court (Berri, 1999).

With the introduction of the book *The Bill James Baseball Abstract* by Bill James, James introduced a new method of analysis for the game of baseball that paved the way for what is more commonly known as sabermetrics (James, 1980). The use of sabermetrics has been further popularized as a result of the continued success of the Oakland Athletics through utilizing the techniques detailed in the book *Moneyball: The Art of Winning an Unfair Game* by Michael Lewis, published in 2003. These unique strategies have allowed the Athletics to exceed the expectations of being a small market team with a comparatively low payroll for Major League

Baseball (Mason and Foster, 2007). Specifically, Wins Above Replacement, WAR, is a particularly motivating statistic behind the increased use of sabermetrics as the measure creates an obvious comparison between each player. By definition, the Wins Above Replacement statistic measures a player's value to his team by determining how many wins he contributes to the team's success as compared with a replacement level player. Multiple methods exist for calculating WAR in baseball, but the statistic remains useful in terms of quantifying player success and contribution to the team in terms of their peers (Major League Baseball, 2021). To successfully use the Wins Above Replacement statistic, an understanding of the role of a replacement level player is necessary as their limited impact creates a basis for the WAR calculations to better comprehend which player is the most valuable to their team. The WAR statistic becomes particularly helpful for determining postseason awards as all players are easily comparable. A replacement level player is defined as a player who contributes a marginal level of production to the team's success but will cost nothing over the league minimum salary to acquire. In baseball, a replacement player is typically defined as a minor league player who could be called up at a moment's notice to replace an injured or recently traded player (Slowinski, 2010). The predecessor to Wins Above Replacement is another statistic that serves a similar role in analyzing player contribution to their team's success. The value over replacement player (VORP) statistic was originally developed for use in Major League Baseball in the early 2000s to measure a player's marginal increase in production as compared to an easily acquired player who can be fairly paid the league minimum (Woolner, 2002). In basketball, the Box Plus Minus (BPM) statistic can similarly be used to measure a player's impact on the game through an integration of all traditional box score information, such as points, rebounds, and assists, and turnovers. However, BPM's usefulness is limited by its inability to quantify a player's playing

time as BPM is a rate statistic based on production per 100 possessions. BPM also struggles to account for less quantifiable data that is not found in the box score, particularly on the defensive end. BPM can be useful for player analysis in basketball as the statistic is used to calculate the value over replacement player statistic in basketball as calculated in Equation (1).

$$\text{VORP} = ([\text{BPM} - (-2.0)] * (\%\text{ of possessions played}) * (\%\text{ of games played})) \quad (1)$$

As shown in Equation (1), the value over replacement player statistic is used to manipulate the factors shown in the equation to calculate a player's individual contribution more accurately. The VORP statistic creates a useful and meaningful method for comparative player analysis that can be applied to both pro basketball in the National Basketball Association (NBA) and college basketball in the National Collegiate Athletic Association (NCAA). A parallel between the VORP and WAR statistics in baseball and basketball is created. Returning to Equation (1), a value of -2.0 is defined to be the level of production for a replacement level player in terms of BPM in the NBA despite the theoretical value for a replacement level player's BPM to be a value of 0. A replacement level player has a BPM value of -2.0 as many below average NBA players spend the majority of the game on the bench as the above average players are likely to accumulate the majority of the available minutes, especially as the season progresses into the playoffs. The next aspect of Equation (1) is the percentage of minutes played which simply expresses the fraction of time that each player is on the court out of the total available minutes throughout the season. This portion of Equation (1) determines the impact a player could have based on the percentage of time the player is on the court. Finally, the last piece of Equation (1) is the percentage of games played that represents the fraction of games played by the player as compared with the team's total number of games. Similarly, the percentage of games played helps to understand the impact that an individual player can have on the game when they are

involved. However, both percentages in Equation (1) can be inaccurate representations of player impact. A player could get substituted into every game within the last moments after the outcome has been decided without having any impact on the game or appear significantly in only a few games. These situations would create a biased Value Over Replacement Player for that player as they would have a higher value despite having less of an impact on the game. These VORP values would be heavily inflated as these players do not have the same impact as a player who consistently produced over the entire season. The value over replacement player statistic does not accurately display who produced over the whole season but who produces based upon the percentage of time on the court. VORP is a rate statistic similar to BPM which fundamentally makes sense as BPM is used to calculate VORP. Building on the value over replacement player statistic, the wins above replacement statistic in basketball is calculated by multiplying VORP by 2.7. This multiplication factor creates a method of basketball analysis to determine each player's contribution to their team's success that is similar to baseball despite the formula for calculating Wins Above Replacement in basketball being vastly different. In both sports, this analysis compares which players are the most integral to their team's success by creating a baseline for player evaluation as all players are compared to the same replacement level player (Myers, 2020).

In this paper, I analyze how to better determine who truly deserves to be selected to these collegiate all-conference teams based on an econometric approach utilizing the Wins Above Replacement statistic. I plan to run a linear regression with multiple independent variables to correct the issues of the Wins Above Replacement statistic and build upon its strengths. I utilize hypothesis testing along with other tests such as the Park test and the White test to ensure that I

am using the correct variables while identifying and correcting any issues that violate the classical assumptions in econometrics such as multicollinearity and heteroskedasticity.

Literature Review

In the current formulation, the Wins Above Replacement statistic has several problems such as the inability to be a known, tangible quantity along with an inability to be easily reproduced, which has led to many different formulations of the statistic. Some of these statistics include openWar and cWAR, and each statistic utilizes a different approach to correct the issues of the WAR statistic. The openWAR statistic was proposed by Benjamin Baumer in an attempt to correct the problems of the WAR statistic by creating a streamlined system of analysis to better understand which players contribute the most to their team's success. The openWar system of analytics utilizes easily accessible data in contrast to WAR, which often uses data that is inaccessible to the public. The system aims for a higher level of transparency in calculations while providing the ability to calculate the statistic at multiple intervals of a given season. The openWar statistic is much closer to a tangible quantity due to better accounting for uncertainty values in its calculation than the WAR statistic (Baumer, 2015). However, openWar has its own problems such as inaccurate predictions and continuing to struggle with defining an accurate replacement level criteria as the statistic uses an arbitrary cutoff in defining who is not a replacement player by excluding the batters with the most plate appearances and the pitchers with the most batters faced with the remaining players being replacement level (Kilanowski, 2020). This arbitrary designation creates a situation where top players getting injured would be labeled as replacement level players which limits the success of the replacement level player definition. Therefore, the openWar statistic does not succeed in correcting all of the WAR statistic's problems as the issue with replacement player definition is still present, which

increases the difficulty to reproduce the calculations necessary to compute openWar. Other research proposes new statistics to overcome the shortcomings of the Wins Above Replacement statistic. One of these statistics is cWAR which analyzes player performance in the Cape Cod Developmental League. This scholarly piece looks to better utilize the wins above replacement technique by more accurately describing what it means to be a replacement level player. The literature describes a replacement level player by being in the league for less than a week, fewer than forty plate appearances for position players, or facing fewer than forty batters for pitchers. By more accurately defining what constitutes a replacement player, any variation of the WAR statistic will have more success in defining player contribution to their respective team (Kilanowski, 2020).

Other Statistical Measures

New statistical measures have been developed in sports outside of baseball and basketball that serve a similar purpose as the Wins Above Replacement statistic. Many of these statistical advances have taken place in the sport of hockey, particularly with the introduction of the Wins over Replacement Player (WORP) statistic (Shea and Baker, 2012) and the Total Hockey Rating (THoR) (Shuckers and Curro, 2013). Specifically, the WORP statistic is formulated to exclusively quantify the contribution of a team's goalie as compared with a replacement level goalie (Shea and Baker, 2012). The first step in calculating the WORP statistic is completed by taking a multiple regression of a team's Goals For (GF) and Goals Against (GA) in explaining a team's total wins. This regression creates a foundation for determining a goalie's importance to their team success as this position has a direct impact on the number of goals scored by the opposing team which directly impacts the success of their team (Shea and Baker, 2012). This work is largely built on Bill James' "Pythagorean Method" in baseball analytics:

$$\text{Win\%} \approx RS^2 / (RS^2 + RA^2) \quad (2)$$

The Pythagorean Method's similar construction to the Pythagorean Theorem leads to its designation. In this scenario, win percentage calculates the number of games won out of the total number of games played while RS denotes runs scored and RA denotes runs allowed (James, 1980). This Pythagorean method from Bill James was adapted by James Cochran and Ron Blackstock to introduce their reimagined win percentage model for use in player analytics in hockey:

$$\text{Win\%} \approx P\ (GP, GF, GA) \quad (3)$$

$$\text{Win\%} \approx GP\ (GF^{1.99063} / (GF^{1.99063} + GA^{1.99063})) \quad (4)$$

The function P in Equation (3) denotes a Pythagorean method with GP representing games played, and GF and GA are goals for and goals against, respectively (Cochran and Blackstock, 2009). Equations (3) and (4) were utilized by Stephen Shea and Christopher Baker to develop the WORP statistic to better analyze goalie performance:

$$WORP \approx P\ (GP_{st}, GF_{st}, GA_{st}) - P(GP_{st}, GF_{st}, GA_{st} + GAS_{rst}) \quad (5)$$

$$WORP \approx ((GP_{st})\ (GF_{st}^{1.99063} / (GF_{st}^{1.99063} + GA_{st}^{1.99063})) - ((GP_{st})\ (GF_{st}^{1.99063} / (GF_{st}^{1.99063} + GA_{st}^{1.99063} + GAS_{st}^{1.99063})) \quad (6)$$

Each statistic represents team *s* and season *t* for goalie *r* in the second term while GP, GF, and GA represent the same values from equations (3) and (4) (Shea and Baker, 2012). The use of the enhanced win percentage model gives the WORP statistic the ability to compare goalies within a single season to determine who had the greatest impact on their team's success while also comparing goalies from any team and season to determine who had the most influence on their team's success in the history of the National Hockey League. However, the statistic remains somewhat limited due to its limited as the statistic analyzes the goaltender position exclusively

and not any other position in hockey. To analyze player performance more broadly within hockey, THoR was developed to compare the statistical contributions of forwards and defensemen more accurately through the use of Markov chains. The foundation of this statistic is derived by using Markov chains to consider every event that occurs within a hockey game and evaluating these events based on the probability of leading to a goal being scored. Furthermore, this statistic considers whether a player began a shift in the offensive or defensive zone, filters out the effects of their teammates and their opponents, and considers whether the particular player is playing at home or on the road. Because of the low scoring rates in an NHL game, the analysis becomes easier to isolate the effects of various game events that directly lead to a goal. Specifically, this statistic analyzes the changes in probability up to twenty seconds after an event occurs as a goal scored outside of the twenty second time frame is insignificantly impacted by the original event. Additionally, a shot's probability of scoring a goal is determined based upon the type of shot taken as well as where the shot occurs on the ice. The statistic only evaluates game play when both teams are at full strength and have all five skaters and a goalie on the ice, which excludes any time spent on the power play, penalty kill, or after pulling the goalie at the end of the game. The even-strength aspect of this statistic is particularly useful as 10 players will be on the floor at all times during a basketball game as no event in basketball compares to a power play in hockey. However, the authors present a model for extending the statistic to uneven player situations such as the power play. To calculate the THoR value, the probabilities of each event that directly creates a goal for the player of their teammates must be accumulated to represent their season long contribution to team's success:

$$\text{THoR} = \text{per event value} * 80 * 82/6 \tag{7}$$

In calculating for the THoR value in Equation (7), the number 80 approximates the number of even strength events that occur in each hockey game. A non-lockout NHL season has 82 games. In order to adjust for players with a smaller sample size, the value is divided by 6. Additionally, this calculation models a player's impact based on theoretically playing the same number of games as compared to an average player whose impact causes few events that create goals, which results in the average player having a THoR rating of zero (Shuckers and Curro, 2013). The formulation of the THoR statistic serves a similar function as statistics such as VORP and WAR as it analyzes individual player contribution over a season compared to a replacement level player. Furthermore, the THoR statistic could be used in conjunction with the WORP statistic to analyze the individual contributions of all positions more accurately within the sport of hockey. The THoR statistic has the unique ability to compare the net goals contributed by each player regardless of how many games played, which makes the statistic particularly useful (Shuckers and Curro, 2013).

Measuring the Success of a Statistic

In addition to the various statistics that attempt to improve on the issues of the wins above replacement statistics, other literature focuses on determining the success of various statistical measures in sports in measuring their intended area of analysis. The accuracy and reliability of a particular statistic depends on the statistic's ability to be discriminating in measuring its intended area, stable enough to provide accurate measurements over time, and must be independent from other statistics. Specifically, a clear differentiation exists between rate statistics and statistics based on total playing time in terms of their characteristics and use in sports analytics. Rate based measures are viewed as less discriminating and more stable while total minutes metrics are believed to be more discriminating and less stable. Rate based statistics

such as Offensive Rating (ORtg), Defensive Rating (DRtg) and BPM analyze player contribution on a per-game or per-minute basis. However, some statistics can more accurately incorporate total playing time into the analysis such as Win Shares, VORP, and WAR. These rate-based statistics are considered more appropriate in estimating a player's skill level whereas the total time metrics more accurately reflect a player's overall seasonal contribution to their respective team which is important for my research (Franks, 2016). Overall, the total minutes metrics are considered more reliable as an increased amount of playing time would suggest a larger contribution to the team's performance. However, total minutes measures alone do not better recognize player ability than any other type of statistic. Because these types of metrics serve two different roles in sports analytics, the statistics should not be directly compared to each other but should be used together to better understand player performance and contribution to their teams. Overall, these statistics are useful in getting a better understanding of who is the most valuable to their team. However, the statistics fall short in terms of the independence criteria for statistical reliability as they fail to accurately quantify the defensive aspect of a basketball game. In the formulation of many rate-based and total minute statistics, there is limited inclusion and influence of defensive statistics (Franks, 2016). Similarly, statistics such as BPM and VORP only account for defensive statistics such as blocks and steals while the BPM statistic equally distributes defensive rebounds throughout the players on the court without regard to their actual production. Other research attempts to quantify the defensive aspects of the game of basketball more accurately. This literature attempts to correct this statistical oversight and counteract the overemphasis of the offensive side of the ball to analyze player performance and contribution to team success more appropriately (Myers, 2020).

Developing Analytics for the Defensive Side of the Ball

To better analyze the defensive aspect of a basketball game, it is an important to understand the roles of the on-ball defender and the players playing help-side defense. Players on defense want to antagonize their offensive counterparts into turning the ball over or taking a low percentage shot unlikely to produce points. A key analytic tool is evaluating the impact of a defender through their ability to contest and prevent easy passes or shots by the offense. Contested passes are considered passes where a defensive player exhibits willful intent to disrupt the offense while being close enough to impede the offensive player's progress. Alternatively, an uncontested pass is defined as willful inaction from the defensive player that results from excessive separation from the offensive player or failing to obstruct the offensive player's movements. By measuring the number of passes contested caused by the defense, the success of the offense can be analyzed in spite of the defense's efforts. Through the literature, contested and uncontested passes are determined to be complements of one another which suggests offensive success can be significantly negatively correlated to the level of contested passes that are pressured by the defense. Therefore, forced turnovers have a more direct and significant impact on limiting offensive performance. The forced turnovers help to outline the relationship of the contested passes and antagonizing offensive performers as contesting passes at a higher rate will often lead to more forced turnovers and a lower field goal percentage for the offensive unit. Despite the potential usefulness of contested passes, any defensive analysis including contested passes is likely too complicated to properly analyze the impact on their opponent's offensive production. This complexity prevents its inclusion in the defensive analysis proposed by Bartholomew and Collier and in statistical analysis in sports in general (Bartholomew and Collier, 2011).

Another piece of literature builds upon defensive statistical inefficiencies by developing new statistical measurements that better analyze the defensive side of the ball. Dean Oliver's *Basketball on Paper: Rules and Tools for Performance Analysis* introduces several statistics to address the lack of defensive analytics. Specifically, he provides a measure that quantifies a defensive player's ability to get stops on an individual basis. A stop occurs whenever a team can prevent their opposition from scoring. Typically, stops are considered on a team basis as five players are needed to successfully get a stop, but Oliver proposes the Individual Stops statistic to quantify how an individual player influences their team's ability to get stops. The five players on the court must work together as a team to get a stop, but the Individual Stops statistic allows for the calculation of which players are the biggest catalyst for effectively getting stops.

Individual Stops = FTO + FFTA/10 + [FM * FMwt * (1- DOR%)] + [DREB * (1- FMwt)] (8)

As shown in Equation (8), FTO is the number of turnovers forced by the defense while FFTA represents the number of free throws missed by a player's opponent. FM quantifies how many shots that a player is directly responsible for causing whereas FMwt is the adjustment for the difficulty of shot caused by the defender as compared to simply getting a rebound from the other team missing a shot.

FMwt = (DFG% *(1-DOR%))/ (DFG% *(1-DOR%) + (1-DFG%) * DOR%) (9)

DOR% represents the percentage of offensive rebounds that the opposing team is able to collect. In contrast, DREB is a team's ability to get a defensive rebound once the opposition misses a shot. DFG% represents the percentage of possessions that the opposition scores on the defense. A player who has a higher number of individual stops will have a larger impact on their team's success and ultimately should have a greater VORP or WAR value, which communicates they are worth more compared to a replacement level player. Each player's defensive success can also

be quantified by the number of scoring possessions allowed, DScPos, where players with a lower DScPos value would be considered a better defender. However, this statistic could create a negative bias toward players who are on the court for a significant number of minutes as they will be responsible for allowing more scoring plays as a result of increased playing time. Therefore, a statistic such as Stop %, which quantifies the percentage of possessions that a defensive player does not allow a basket or commit a foul, is more likely to accurately reflect the true defensive talent of a player as the statistic communicates a player's individual defensive influence on the game. DRtg is another defensive statistic that attempts to quantify an individual player's impact on the defensive end of a game. However, this calculation is largely based on their team's defensive rating which could create a negative influence that is difficult to overcome if a player is an above average defender playing on a team that does not defend well.(Oliver, 2004).

Determining an Individual's Contribution in a Team Sport

Another key determination in accurately assessing a player's individual contribution is filtering out the contributions of their teammates. This consideration is important for differentiating the best players from the players who are surrounded by the most talented team. Often, losing teams fail to accumulate a large enough number of positive statistics or end up creating too many negative ones. However, a player's statistical value to their team should be independent of the teammates around them, and they should be evaluated consistently from team to team. Therefore, a focus on the player's individual statistical output is necessary to evaluate their true worth to their team (Berri, 2019). Specifically, several statistics function to remove the impact of a player's teammates to assess individual impact more accurately. The WORP statistic was mentioned earlier, but another version of the statistic analyzes a goaltender's performance

by removing the impact of the goaltender's teammates and opponents. This individualized version of WORP is Team Independent WORP, or TIWORP. This variable can be calculated with the following equation:

$$TIWORP = b_t * (minutes_{rst} / minutes_t) * SOG_t * (SV\%_t - SV\%_{rst}) \quad (10)$$

The b_t values are shown in Table 1, which are found by the taking a multiple regression of the goals for and goals against variables to determine the number of wins created by a goalie while the $minutes_{rst}$ and $SV\%_{rst}$ are the minutes played and average save percentage, respectively, for goalie *r* on team *s* in season *t*. Similarly, $minutes_t$, SOG_t, $SV\%_t$ are total minutes, total shots on goal, and average save percentage, respectively, for goalies in season *t*. The Team Independent WORP value allows for adjusting a goalie's WORP value by removing the impact from his respective team. WORP is often biased by the number of shots that a goalie sees as more shots faced leads to a higher WORP value, so the TIWORP value is able to correct this bias from the number of shots faced to better evaluate goaltenders (Shea and Baker, 2012).

Table 1: Beta Values Used to Calculate TIWORP from 1982- 2010

Season	b value	Season	b value
1982-83	-0.1380	1996-97	-0.1607
1983-84	-0.1362	1997-98	-0.1790
1984-85	-0.1201	1998-99	-0.1660
1985-86	-0.1347	1999-2000	-0.1562
1986-87	-0.1598	2000-01	-0.1719
1987-88	-0.1510	2001-02	-0.1459
1988-89	-0.1362	2002-03	-0.1724
1989-90	-0.1441	2003-04*	-0.1691
1990-91	-0.1579	2005-06*	-0.1724
1991-92	-0.1523	2006-07	-0.1823
1992-93	-0.1679	2007-08	-0.1785
1993-94	-0.1303	2008-09	-0.1736
1994-95	-0.1376	2009-10	-0.1782
1995-96	-0.1644		

*2004-05 NHL season did not occur because of the lockout

Additionally, the THoR statistic eliminates the impact of teammates when evaluating individual player performance. By utilizing a probability-based approach and defining a period of influence of twenty seconds, the statistic is able to limit the impact of teammates and opponents which makes the statistic's analysis of individual performance more accurate (Shuckers and Curro, 2013).

Measuring a Teammate's Impact

The impact that teammates have on each other, both on and off the playing field, is another important consideration for evaluating players. This impact can be largely attributed to knowledge spillover and peer pressure effects. Knowledge spillover occurs when teammates and competitors learn from each other and improve their skills as a result of ongoing practice and competition while peer pressure effects can cause a player to raise their performance level to meet the expectations of the team or to match the efforts given by their teammates. These effects make it even more necessary to understand a player's impact without their teammates to determine what their true level of performance contributes to the team's success. The research from Molodchik, et al (2021) investigated 5000 players on more than 230 teams and discovered that players on stronger teams perform better than expected. However, stronger teams have a diminishing rate of marginal improvement in affecting an individual's performance. Also, a roster with players of varying skill levels is motivating for the less talented players to learn from the experienced players and improve their skills quickly. A low percentage of new players joining the team is also better for player performance as the need for integrating new players and learning how to play with them is limited. Fewer new players require less time helping acclimate the new players to the team, so the individual can focus on their own performance. However, a higher quality new player is more likely to improve a player's individual performance as they are

able to positively impact the team overall without limiting the team's learning or growth (Molodchik, et al, 2021).

Considering an Individual's Impact by Their Team's Performance in Their Absence

Another approach for determining a player's individual contribution is to analyze the team's performance when the individual is unable to participate. This situation can arise when a player is injured or being rested at the end of the season as the team has already clinched a playoff berth. In soccer specifically, a player can also not be in the lineup when they are representing their home nation in a tournament such as the African Cup of Nations or in World Cup qualifying. Typically, only the elite players will be given the honor of representing their country, so these select few players will know far in advance when they will be playing for their country. This type of absence allows the player's team to plan ahead for the time when the player is away from the club. However, injuries cannot be predicted, so teams are forced to react when injuries happen to replace the injured player's production. This piece of literature uses a multiple linear regression to analyze the data concerning these soccer players who participate in the African Cup of Nations from the top European soccer leagues, such as the Premier League and the Bundesliga League, to determine the significance of losing these players for an extended amount of time. The research explains this significance through player quality, team quality, and player salary. By running this regression, the author of this piece concludes that these teams experience no significant loss from these players participating in the African Cup of Nations as many of these soccer clubs have great depth which allows them to easily replace the player competing in the tournament. Therefore, this method for evaluating player contribution cannot be considered useful based on the findings of Perez's article. However, this piece only considers players absent due to national team appearances and not due to injury, so this area could need

more investigation (Perez, 2021). In contrast with top league club teams in soccer that have enough depth to combat this scenario, basketball teams are limited to a fewer number of players on the roster. In the NBA, each team can only have fifteen players on the roster at a time while being limited by the salary cap set by the NBA to maintain a competitive balance across the league. The salary cap limits how much talent can be brought to the roster, especially if the team has superstar players who are on max contracts (Engler, 2011). Similarly, collegiate teams can only have a certain number of players on the roster as these college teams are largely inhibited by the number of scholarships that a school can offer, which the NCAA dictates. This scenario creates a possible depth issue for these basketball teams as their rosters will have less talent at one time, which can limit a team's ability to combat losing a key player for an extended period of time as compared with a top soccer league in Europe (Engler, 2011).

Performance in Late Game Situations

When determining a player's contribution to their team's success, it is also important to consider how that specific player performed in different periods of the game, particularly whether their performance improves as the game progresses toward the end of regulation and into overtime. As games advance, external and internal pressures can build and impact a player's performance specifically in critical moments which are defined as the last five minutes of regulation and any overtime periods (Gomez, et al, 2015). Furthermore, another piece defined critical moments as the last two to three minutes of a game with a difference in the game score of less than ten points (Ferreira, Volossovitch, and Sampaio, 2014). In comparison, other research defines a critical moment to be the last five minutes of regulation and overtime where the score is within less than three points (Annis, 2006). These pressures can cause a negative disruption in how a player is able to execute basketball actions such as dribbling, passing, and ball-handling.

A player is particularly vulnerable to deteriorations in their performance level and decision making as stress and pressure build late in the game, which often results in turnovers and poor shot selection. Because of increased stress, accumulation of fatigue, increased fouling, and a higher frequency of timeouts that occur during the critical moments of a game, coaches must prepare appropriate strategies that cater to these critical moments. Through this piece of literature, Gomez runs a binomial logistic regression that estimates the regression weights along with the impact of a variety of situational variables and game related statistics. The regression leads to the conclusion that many box score statistics are less important in the overtime period as coaches adjust their strategies to protect the lead and combat player fatigue. Winning teams are more likely to have better production in terms of shooting, rebounding, blocks, and steals while being less affected by the pressure of the situation or the pressure applied by the other team. On the contrary, losing teams are much more affected by the situation and their opponent as they commit more turnovers, miss more three-point attempts, and commit more fouls. The regression argues that the only significant variables in the overtime period were successful three-point field goals, successful free throws, and fouls committed. The regression also posits that home field advantage does not exist in the overtime period as the home team is more likely to falter in the overtime period as the visiting team has already adjusted to the distractions imposed by the fans and the situation (Gomez, et al, 2015).

A similar pressure buildup occurs in hockey as a more direct impact is exerted by an individual player's performance as a hockey game moves into overtime or shootout as the game remains tied. During the 2018-2019 season, the National Hockey League altered the overtime format from each team having 4 skaters and a goalie to playing with 3 skaters and a goalie. In the shootout period, players from the two teams trade opportunities to score a goal against the other

team's goalie without any other defenders impeding the attempt. Naturally, as the game progresses, each player will have a greater impact on the outcome of the game as the increased space on the ice allows increased opportunity to make a play with fewer players on the ice. These game situations create an opportunity for each player to exert a greater impact on the game's outcome. The overtime period increases the pressure on players in a similar fashion as overtime in basketball; however, the decreased number of players leads to an increased amount of attention and scrutiny on each player, which is unique to hockey (Hoffman, et al, 2021).

Foundations of Econometrics

When utilizing an econometric approach, the classical econometric assumptions must be upheld to validate the model being used. Heteroskedasticity and multicollinearity are two key issues that violate these classical assumptions. Heteroskedasticity occurs when the variance of the error terms varies with the independent variables, which violates the classical assumption that the population errors are constant; Multicollinearity occurs when one variable has a direct linear relationship with another variable in its formulation which causes a high level of correlation between the variables. Multicollinearity can violate the classical assumption that variables are not correlated with each other. Perfect multicollinearity is a particular issue for violating this assumption as perfect multicollinearity is the more severe case of multicollinearity and causes improper conclusions to be made as two variables have an exact linear relationship (Studenmund, 2016). Other literature investigates how to best overcome the issues of heteroskedasticity and multicollinearity to properly model the research topic at hand, which will be of key importance to my research. In sports, large amounts of data are readily available for analysis due to recent advances in technologies such as Global Positioning Systems (GPS) (Weaving, et al, 2019). However, dealing with this massive amount of data can be daunting, and

this data often has perfect multicollinearity as many of the statistics can be highly correlated to each other. Weaving and his colleagues suggest an approach using Leave One Variable Out, Partial Least Squares Analysis (LOVO-PLSCA) that is designed to correct the issue of multicollinearity and highlight the most influential variables in determining the fitness level of athletes in training by comparing the results of each of the LOVO-PLSCA regressions. The LOVO-PLSCA method is immune to the effects of perfect multicollinearity to discover which variables are the most influential in modeling the output. The LOVO-PLSCA method helps the multiple linear regression avoid the problem of multicollinearity as the method excludes the use of variables that have direct linear relationships. This method also utilizes the correct explanatory variables, which increases the accuracy of the regression in modeling the fitness scenario described in the literature but also in all sports performance scenarios. This study contends that an analysis using the Leaving One Variable Out, Partial Least Square regression technique is useful in running a more effective multiple linear regression so that the most influential explanatory variables are utilized in the model (Weaving, et al, 2019).

Another piece of literature attempts to combat multicollinearity by using structure coefficients to scale the beta values more appropriately for the regression equation. Beta weights alone can often lack reliability as they create the bouncing beta problem which occurs as multicollinearity arises. The bouncing beta problem causes the beta values from the multiple linear regression to be inaccurate representations as the beta values are pulled in the same direction (Henson, 2002). This issue necessitates the use of structure coefficients alongside the beta coefficients. The structure coefficients are found through a bivariate correlation between a single observed predictor variable and the output variable. This definition avoids the issue of perfect multicollinearity as only one dependent variable is regressed against the output, so the

variables cannot be defined in terms of each other or have a direct linear relationship. However, the structure coefficients are not independent of the other independent variables' effects despite not being directly influenced by the relationships between each independent variable. More specifically, the structure coefficients display the extent of the dependent variable's variance that can be explained by a single independent variable, which suggests that the sum of the squared structure coefficients together would highlight the breakdown of each independent variable's impact on the variance of the dependent variable. However, a consideration of the beta values and the structure coefficients together is important for modeling the situation more accurately to avoid inaccurately and unnecessarily increasing the beta values in attempt to model the relationship (Yeatts, et al, 2017).

Some literature also tackles the issue of heteroskedasticity when utilizing multiple regression models. Heteroskedasticity occurs when the variance of the population errors is not constant; this issue causes a violation of the classical econometric assumptions which will cause the slope of the regression line to be inaccurate in estimating the relationship between the independent variables and the dependent variable (Studenmund, 2017). In the NBA, heteroskedasticity occurs as the strength of the team varies with an increase in the number of games. Their ranking compared to the rest of the league fluctuates throughout the season which creates heteroskedasticity as their strength is not constant as the year goes on. This piece relaxes the constraint of a constant variance error to model team strength with $S_{ik} = \beta + e_{ik}$ where β is the constant component of team strength, and the error team is normally distributed for team k in year i. This definition allows for a higher level for the variance which implies a more volatile performance level for each NBA team. This high volatility can be negatively influenced by issues in scheduling such as playing games on back-to-back nights and benefitting from playing

on their home court in front of the team's fans. This model evaluates the NBA performance level over multiple seasons to determine if heteroskedasticity influences the estimation of a team's strength and their ranking relative to the rest of the teams in the NBA. However, the analysis finds only weak evidence of heteroskedasticity as home court advantage only causes an increase of 2.7 points whereas playing on back-to-back nights costs teams 1.8 points. The study also determines that these point variances happen to successful and unsuccessful teams alike. Additionally, some volatility occurs due to injuries during the season, but these fluctuations are much less predictable and not as easily fit into the model (Manner, 2016).

Other Econometric Models

Other research aims for a similar goal using similar methods as my thesis, which is a direct measurement of player productivity to determine which players are truly the most pivotal to their team's success. The most significant piece of work in this area presents an econometric model that connects player production to wins created based upon the player's accumulation of certain key statistics, including offensive rebounds and assists. This model creates a more accurate representation of player value to replace outdated awards, such as the IBM award, that claim to be more objective than they actually were; similar models represent this measure as a summation of a player's positive statistics minus the summation of their negative ones, even if basketball does not always happen in such a binary fashion. This literature runs a regression of defensive turnovers created, defensive rebounds, defensive points allowed, number of turnovers, points per possessions, free throws attempted, free throws made, and offensive rebounds to model player production. According to this research, offensive rebounds are considered to have the greatest impact with three-point field goals made being a close second. Additionally, ball handling has extreme importance as turnovers by either team have a great impact on the outcome

of all games. Furthermore, efficient shooting is more important than scoring a higher total number of points, particularly if a large number of shots is needed to achieve the high point total. Therefore, a player's ability to produce more wins is largely dependent on the ability to acquire and keep possession of the ball while consistently making field goals, which largely translates to an accumulation of rebounds, avoiding turnovers, and shooting efficiency. To properly account for a player's production, the literature analyzes per-minute player production, per-minute team tempo factor, and per-minute team defense by accumulating each of the variables included in the regression.

Per-minute production (PM) = accumulation of each statistic/ Total minutes played (11)

This factor attempts to quantify how productive a player is per minute, which will be easily comparable between players.

Per-minute team tempo factor (TF) = (Team FGA * -0.023 + Team FTA * -0.009)/ Total minutes played (12)

Team FGA represents team field goal attempts while Team FTA represents team free throw attempts. This aspect looks at how many possessions a player's team has per minute to determine how production is affected by their team's pace.

Per-minute team defense (TDF) = accumulation of defensive statistic/ Total minutes played (13)

This factor specifically looks at summing the defensive statistics to accurately determining the player's impact on the defensive side of the ball.

Wins produced = (PM + TF + TDF – PA+ TA) * total minutes played (14)

In Equation (14), PA represents average per minute production at position, which allows the function to determine how a player's per minute production compares to the average production for their position. Also in Equation (14), TA represents an average player's per-minute

production, which allows the win production function to fully evaluate all player's contributions in comparison with one another. At the conclusion of this piece of literature, an analysis of the 1997-1998 NBA season applies the findings of the research to determine which player should have been should have been selected as the Most Valuable Player. Surprisingly, this analysis results in Dennis Rodman, not Michael Jordan and Karl Malone who were the top MVP candidates from that season, being highlighted as the most valuable player to his team due to the impact that his offensive rebounding prowess provided (Berri, 1999).

My Original Contribution:

To properly build a model accurately evaluating player performance in collegiate basketball, I utilize a population regression line, abbreviated as PRL, to predict a player's contribution to their team's success based on several key statistics. The variables included in the original specification of the model are shown in Table 2 below.

Table 2: Description of Variables

Variable	**Abbreviation**	**Description**
Turnovers Per 100 Possessions	TO / 100 Poss	Number of Turnovers committed in 100 possessions
Individual Stops	Individual Stops	Number of Stops that a player individually contributes
Free Throw Percentage	FT%	Success rate in making free throws
Offensive Rating	ORtg	The number of points scored per 100 offensive possessions
Defensive Rating	DRtg	The number of points allowed per 100 defensive possessions
Defensive Rebound Percentage	DREB%	The percentage of defensive rebounds that a player secures
Stop Percentage	Stop%	The percentage of possessions that a player prevents an opponent from scoring
Effective Field Goal Percentage	eFG%	Success rate in making shots while weighting shots according to their point value

These variables allow me to fully analyze player performance with an increased emphasis on the defensive aspect of the game through the inclusion of the Individual Stops and Stop% variables. I utilize hypothesis testing to investigate how the success of these variables to output a player's WAR value. The players with the highest WAR values will be selected to the all-conference teams for their respective season with the player with the highest WAR value being selected as the conference player of the year. The initial specification of the PRL is shown in Equation (15) below.

$$\text{WAR} = \beta 0 + \beta 1 * \text{TO/100 Poss} + \beta 2 * \text{Individual Stops} + \beta 3 * \text{FT\%} + \beta 4 * \text{ORtg} + \beta 5 * \text{DRtg} + \beta 6 * \text{DREB\%} + \beta 7 * \text{Stop\%} + \beta 8 * \text{eFG\%} \quad (15)$$

The individual stops statistic is not typically considered in basketball analytics, so several aspects of this statistic had to be estimated within my analysis. The Individual Stops statistic is shown in Equation (8) earlier in the paper. An important aspect of calculating the Individual Stops statistic is calculating how many turnovers a player individually causes. This calculation for turnovers caused is estimated in my analysis as shown in Equation (16) below:

$$\text{FTO} = (\text{STL\%} + \text{BLK\%}) * \text{Team FTO} \quad (16)$$

To better represent a player's individual turnovers forced in Equation (16), STL% represents the percentage of possessions that a player steals the basketball from the other team. Likewise, BLK% represents the percentage of possessions that a player blocks an opponent's shot. Team FTO represents the total number of turnovers that a team causes for their opponent within a season. Next in Equation (8), FFTA is represented by Equation (17) below.

$$\text{FFTA} = (\text{OPP total missed FT}) * (\text{minute played \%}) / 10 \quad (17)$$

The FFTA value approximates the number of missed free throws that a player is directly responsible based on the total free throws missed by opposing teams and the player's time on the

court. The FFTA values is divided by 10 to distribute the impact of the missed impact between an estimated 10 rotation players. This statistic is also not typically recorded in basketball analytics, so the value of FFTA must be estimated to determine the Individual Stops variable. The next aspect of Equation (8) that needs to be understood is forced misses, which is represented by FM.

$$FM = \text{Total OFGM} * (\text{minutes played \%}) / 5 \quad (18)$$

Total OFGM represents the total number of shots that missed by the opponent. This figure is adjusted by the minutes played percentage and divided by five to isolate the number of misses that a singular player causes among the five players on the court at one time. Forced miss weight, represented FMwt, was shown earlier in the paper in Equation (9) to estimate the percentage of shots missed as a result of a player's defensive impact that increases the opponent's difficulty of shots taken. The statistics used to calculate FMwt are also useful in calculating Individual Stops. Returning to Equation (15), a higher ORtg would suggest higher offensive prowess while a higher DRtg would highlight a player as a defensive liability. The Stop% statistic determines how many misses a player helped cause by determining the number of possessions that a player was on the court. Much like Individual Stops, Stop% is not a widely used statistic, which causes the measure is not heavily reported or calculated. In this paper, the statistic is estimated as shown in Equation (19) below.

$$\text{Stop\%} = \text{Total OFGM} * (\text{minutes played \%}) / \text{Total Opp Poss} \quad (19)$$

The final variable from Equation (15) is eFG% variable, which is calculated with Equation (20) below.

$$\text{eFG\%} = (\text{Field Goals} + (0.5 + \text{3pt Field Goals})) / \text{Field Goal Attempts} \quad (20)$$

This variable weighs a player's shooting to more accurately reflect the point value of the shots taken. This proportion allows shooting a 2-point field goal to have the same impact on field goal percentage as a 3-point field goal by weighing 3-point field goals more to reflect being worth more points.

This proposed model, as shown in Equation (15), utilizes variables that are readily available statistics in basketball analytics with the exception of the Individual Stops and Stop% variables. The integration of these under-utilized variables allow for an increased emphasis on the defensive side of the game as their use in basketball analytics is uncommon given their exclusivity to Dean Oliver's book *Basketball on Paper: Rules and Tools for Performance Analysis* as exhibited earlier in the Literature Review.

Regressions and Analysis

My analysis examines ten years of ASUN conference player data as found on Basketball Reference to construct a regression-based model that determines the true impacts of the variables included in the original model in Equation (15). In conjunction with the WAR variable that was outlined in the literature review, these variables are used to create a more accurate and effective model to measure the performances of the top players in the ASUN to better select players to the all-conference teams with an increased emphasis on the defensive aspects of basketball. In creating my model, I established a restriction that players must play more than half of their team's available minutes so that only the truly influential players would be discussed within the model's analysis. An inclusion of players who do not meet this criterion would decrease the model's ability to analyze player performance. Many players only play when the game's outcome is already decided, so it would be inappropriate to include these players in the analysis as their influences would be inflated and inaccurate based on their limited court time. In a typical

season, a team usually has three to six players who played more than 50% of their team's minutes. With eight to twelve teams in the league each year, a typical season would have approximately 40-50 players in consideration for all-conference selection by the model. In total, 440 players were analyzed over a ten-year period. Because my analysis examines ten years of data, I believe that six years would be an appropriate number of seasons to begin testing if each of the proposed variables were significant in determining a player's WAR value. This determination is made through hypothesis testing by comparing if the variable's probability to a given level of alpha. A variable is significant at a specific level of alpha if the variable's probability value is less than that level of alpha such that a lower level of alpha indicates a higher level of significance. A given level of alpha indicates the level of influence that a variable has in calculating the output value of the model with a lower level of alpha communicating a higher level of significance. Another consideration for the explanatory power of the model is the R^2 value which communicates the model's ability to determine how consistently the output from the independent variables matches the values used as the dependent variable in the regression. The R^2 value is a measure of goodness of fit for the model. Additionally, the adjusted R^2 value alters the R^2 value to correct for multiple independent variables. The results for the first 6 years of data are shown in Table 3.

Table 3: Model Iteration 1 with 6 years of data

6 years of data	Adj R^2 = 0.674651	N=242	F value=2.73E-54
Variable	**Coefficient**	**Probability value (P-value)**	**Alpha level of significance**
TO/100 poss	0.097738	0.630	Insignificant
Indiv Stops	0.117399	7E-06	α= <1%
FT%	5.294986	0.019	α= 2%
ORtg	0.155079	5E-07	α= <1%
DRtg	-0.273771	5E-14	α= <1%
DREB%	-0.26174	0.958	Insignificant
Stop%	28.41361	2E-09	α= <1%
eFG%	23.45982	5E-06	α= <1%

It is important to note that scientific notation will be represented with an E as shown in Table 3 where 2.73E-54 represents 2.73×10^{-54}. Based on the results in Table 3 after 6 years, my model has an adjusted R^2 =0.674651 which suggests that the data and its output fits within the expected value for model 67.4651% of the time. To determine the joint significance of my model, I utilize the F test.

H_o: $\beta1= \beta2= \beta3 = \beta4= \beta5= \beta6= \beta7 = \beta8= 0$

H_a: At least one of $\beta1, \beta2, \beta3, \beta4, \beta5, \beta6, \beta7, \beta8 \neq 0$

Because 2.73E-54< α= <1%, I reject the null hypothesis and accept the alternative hypothesis at α = <1%. This conclusion means that the independent variables are jointly significant at α= <1%. Therefore, at least one of the slopes does not equal 0. This significance communicates that the variables work well together to explain the output of the model. A similar hypothesis test is applied to each of the independent variables.

Null hypothesis (H_o): $\beta_i = 0$

Alternative Hypothesis (H_a): $\beta_i \neq 0$

Each hypothesis test is applied to determine each variable's significance as i represents which variable is being tested. In this analysis, i can range from one to nine. To test for the level of the

significance, various levels of alpha are applied to the p-value of each variable as the variable becomes significant at the level of alpha where the p-value has a smaller value than the specified level of alpha. A lower alpha level suggests a higher level of significance for the variable in modeling player performance accurately. The highest acceptable level of alpha that is significant is $\alpha = 20\%$, and variables whose p-values are larger than $\alpha = 20\%$ are considered insignificant and should be removed from the model as that variable would inaccurately model player performance. The results of these hypothesis tests are shown in Table 3.

After regressing 6 years of data, the variables Individual Stops, ORtg, DRtg, Stop%, and eFG% all had p-values that were significant at $\alpha = <1\%$, which suggests that these variables are highly influential in determining a player's WAR value for my model. The FT% variable was significant at the $\alpha = 2\%$ which is still quite significant in determining the WAR value for players but less significant than the variables significant at $\alpha = <1\%$. However, the variables TO/ 100 possessions and DREB% were highly insignificant which makes the variables inconsequential in calculating a player's WAR value in the model as the variables have a p-values of 0.630 and 0.956, respectively, so I am going to remove the variables from the model at least for now. These methods of hypothesis testing will be key for my analysis in determining the significance of each variable independent of the other variables while also testing the joint significance of the variables. The theory of these tests will be applied to later iterations of the PRL, but the tests will not be explicitly shown in the analysis as only their results will be discussed. After removing TO/ 100 Poss and DREB%, the new PRL becomes:

$$WAR = \beta 0 + \beta 1 * \text{Individual Stops} + \beta 2 * FT\% + \beta 3 * ORtg + \beta 4 * DRtg + \beta 6 * Stop\% + \beta 7 * eFG\% \quad (21)$$

After removing the insignificant variables, I included additional variables in the model to uphold the model's explanatory power and ensure that I did not have too few variables. Therefore, I added variables for a player's total turnovers in a season, represented by TO, a player's total assists for the season, abbreviated as AST, and their total offensive rebounds for the season, shown as OREB. After adding in the new variables, the PRL becomes:

$$WAR = \beta 0 + \beta 1 * \text{Individual Stops} + \beta 2 * FT\% + \beta 3 * ORtg + \beta 4 * DRtg + \beta 5 * Stop\% + \beta 6 * eFG\% + \beta 7 * TO + \beta 8 * AST + \beta 9 * OREB \quad (22)$$

I believed that adding these were influential statistics that determined a player's impact on the game, so I ran another the regression with 6 years of data to determine the impact of these additional variables on the model with the results shown in Table 4.

Table 4: Model Iteration 2 with 6 years of data

6 years of data	R^2 =0.687499	N=242	F value = 8.01E-57
Variable	**Coefficient**	**P-value**	**Alpha level of significance**
Indiv Stops	0.118478	1.131E-05	α= <1%
FT%	5.961289	0.013248	α= 2%
ORtg	0.108534	0.000504	α= <1%
DRtg	-0.29525	6.26E-17	α= <1%
Stop%	18.43162	1.89E-05	α= <1%
eFG%	28.63112	1.57E-08	α= <1%
TO	-0.01817	0.236929	Insignificant
AST	0.031067	0.000176	α= <1%
OREB	0.004495	0.727585	Insignificant

After running this regression, the adjusted R^2 value increased to 0.687499, which suggests that the addition of the turnovers, assists, and offensive rebounds variables improved the validity of the model. The model now has a 68.7499% fit for predicting the model's output which is in an increase in value from the previous value of 67.4651%. Additionally, the F value decreased to 8.01E-57 which suggests that the variables are now more jointly significant than the previous iteration of the model given the increased distance from the α= <1% value for the F test. The

variables Individual Stops, ORtg, DRtg, Stop%, and eFG% remain individually significant at an alpha value of less than 1% which suggests that these variables are highly influential in determining a player's WAR value in my model. The FT% variable remained significant at the alpha equal to 2% which is remains significant in determining the WAR value for players but less significant than the first variables significant at less than 1%. The newly introduced TO variable was not significant at any levels of alpha as the variable has a p-value of 0.236929 which suggests that the variable has limited impact on player impact in this iteration of the model. Despite the lack of significance for the TO variable, I left the variable in the model for the time being given the high theoretical relevance of turnovers in determining a player's impact on the game. The AST variable had a p-value of 0.000176 which means that the variable is highly significant in explaining player performance at an alpha level of less than 1%. As the OREB variable is introduced, the variable has a p-value of 0.727585. This high p-value suggests that the variable has minimal impact on player performance. Because of a lacking impact and statistical insignificance for the OREB variable, I took OREB out of the model. After the OREB variable was removed, the PRL becomes:

$$\text{WAR} = \beta 0 + \beta 1 * \text{Individual Stops} + \beta 2 * \text{FT\%} + \beta 3 * \text{ORtg} + \beta 4 * \text{DRtg} + \beta 5 * \text{Stop\%} + \beta 6 * \text{eFG\%} + \beta 7 * \text{TO} + \beta 8 * \text{AST} \quad (23)$$

After adapting the PRL as shown in Equation (23), I added an additional two years of data to my regression to supplement the analysis. The results for the eight-year regression are shown in Table 5.

Table 5: Model Iteration 3 with 8 years of data

8 years of data	Adj R^2 = 0.613932	N=336	F value = 9.33E-65
Variables	**Coefficient**	**P-value**	**Alpha Level of Significance**
Individual Stops	0.177026	6.96E-14	α= <1%
FT%	3.26939	0.108127	α= 11%
ORtg	0.135131	1.18E-05	α= <1%
DRtg	0.006679	0.105506	α= 11%
Stop%	23.33236	1.19E-07	α= <1%
eFG%	30.48836	8.74E-10	α= <1%
TO	-0.02823	.045894	α= 5%
AST	0.026802	8.64E-05	α= <1%

After 8 seasons of data, the model has an adjusted R^2 =0.613932 which suggests that the model has a 61.3932% ability to forecast the relationship between the input variables to output the output variables. This decrease is a bit concerning as the fit of the model has decreased by approximately seven percent, but the additional data is meant to help smooth out the data's variability throughout the eight seasons to determine which players are truly the most influential and most worthy of an all-conference team selection. In contrast, the F value of this iteration is 9.33E-65 which is a smaller value than the previous version of the model. This decrease suggests that the variables in the model are more jointly significant than previously with the same α= <1%. The variables for Individual Stops, ORtg, Stop%, eFG% and AST are significant at α= <1%, which suggests these variables are highly influential in determining player performance and impactful on team's success. The TO variable was significant at an α= 5%, which suggests that the variable is influential in determining a player's performance level but less impactful than the variables that were significant at an α <=1%. The significance of the TO variable confirms my decision to leave the variable in the model given its theoretical relevance as the variable is now statistically significant as well. The variables for Free Throw percentage and DRtg were only significant at an alpha level of 11% which suggests some influence on player performance

but certainly less impact than the variables significant at both the α <=1% and α = 5%. This trend is somewhat concerning as the variables have become much less significant than the previous iterations of the model, so the downward trend of these variables is worth monitoring. Because I did not remove or add any variables, the PRL remains the same as shown in Equation (23), so I proceed in adding the final two years of data to complete my analysis of the ASUN conference data. Table 6 below shows the results of the regression for the ten-year data set.

Table 6: Model Iteration 4 with 10 years of data

10 years of data	R^2 = 0.610481	N=440	F value = 4.44E-85
Variables	**Coefficient**	**P-value**	**Alpha level of significance**
Individual Stops	.157231	7.98E-18	α= <1%
FT%	.487441	0.331557	Insignificant
ORtg	.188754	1.95E-12	α= <1%
DRtg	.004863	0.243251	Insignificant
Stop%	21.52845	9.01E-08	α= <1%
eFG%	24.64479	3.04E-08	α= <1%
TO	-0.0112	.364097	Insignificant
AST	0.02245	0.000185	α= <1%

After 10 years of data, the adjusted R^2 value has decreased even further to 0.610481, which communicates the model has a 61.0481% ability to accurately utilize the input variables to model player performance through the WAR statistic. This subsequent decrease is a bit concerning as the model's predictive power has decreased modestly within the last two iterations of the model. However, I attribute these decreases to increasing the data set which was expanded to avoid issues of variability from a sample size that is too small and does not reflect the variances in performance between players across the seasons within this ten-year period. The F value has further decreased to a value of 4.44E-85 which signifies that the variables are even more jointly significant than the previous versions of the model. The Individual Stops, ORtg, Stop%, eFG%, and AST variables remain significant at α= <1%, which suggests that these variables remain

highly influential in modeling player performance within the model. I noticed a marginal increase in the p-value for the AST variable. This change suggests that the variable's influence has slightly decreased, which is a concerning trend that merits further attention. The FT%, DRtg, and TO variables have all become statistically insignificant as the variables have higher p-values than any acceptable levels of alpha that suggest the variables as having a significant impact on player performance within the model. Because of this lacking impact, I considered removing these variables from the model. Before removing the insignificant variables from the model, I attempted to systematically remove these variables to determine if any of them actually had an individually significant impact on the model despite an insignificant p-value. The result of these removals is shown in Table 7.

Table 7: Removing FT%, TO, DRtg, Sequentially

	All 3 variables still in the model	After removing FT% - DRtg and TO still in	After removing TO, only DRtG remains	All 3 removed
Adj R^2	0.610481	0.610531	0.610793	0.610435
P-values				
FT%	0.331557	N/A	N/A	N/A
TO	0.243251	0.400451	N/A	N/A
DRtg	0.364097	0.244973	0.237408	N/A

As shown in Table 7 above, removing these variables from the model was the correct decision as the variables never became statistically significant even as the variables were removed one at a time despite a fluctuating adjusted R^2 value. Once these variables were removed, the updated PRL is shown in Equation (24).

$$WAR = \beta 0 + \beta 1 * \text{Individual Stops} + \beta 2 * \text{ORtg} + \beta 3 * \text{Stop\%} + \beta 4 * \text{eFG\%} + \beta 5 * \text{AST} \quad (24)$$

Table 8 displays the results of the latest regression after removing the insignificant variables.

Table 8: Model Iteration 5 with 10 years of data

10 years of data	Adj R^2 =0.610435	N=440	F-value = 1.4E-87
Variables	**Coefficients**	**P-values**	**Alpha Level of Significance**
Individual Stops	0.148553	2.6E-19	α= <1%
ORtg	0.204173	9.11E-18	α= <1%
Stop%	21.17773	5.85E-08	α= <1%
eFG%	23.15632	6.15E-08	α= <1%
AST	0.019477	6.06E-05	α= <1%

Even though removing these variables from the model was the correct theoretical decision, I was concerned with the model's ability to correctly forecast player production with so few variables which creates an inaccurate representation of each statistic's ability to impact a player's contribution to their team's success. This concern was amplified by another decrease in adjusted R^2, which communicates a problem in the model's predicting power. The variables that remain are all significant at α= <1%, which suggests that all of these statistics are highly influential in determining player performance. However, these influences could be overstated given the current lack of variables in the model. After this consideration, I began considering other variables that could be defining measures of success within college basketball to include in the model. Because of the theoretical influence of defensive rating and free throw percentage in determining a player's influence on his team's success, I decided to reexamine my data to look for any incorrectly entered data that would hinder the potential for these variables to be significant. I was extremely surprised to actually find two instances of data entry errors: a DRtg was intended to be 95.8 but was listed as 958 while another player incorrectly had a zero percent free throw percentage. After reincluding the DRtg and FT% variables, the updated PRL appears in Equation (25) below.

$$WAR = \beta 0 + \beta 1 * \text{Individual Stops} + \beta 2 * \text{ORtg} + \beta 3 * \text{Stop\%} + \beta 4 * \text{eFG\%} + \beta 5 * \text{AST} + \beta 6 * \text{DRtg} + \beta 7 * \text{FT\%} \quad (25)$$

I ran another regression using the PRL shown in Equation (25) to determine the effects of this discovery as shown in Table 9 below.

Table 9: Model Iteration 6 with 10 years of data

10 years of data	Adj R^2 =0.714458	N=440	F value = 7.7E-115
Variables	**Coefficient**	**P-value**	**Alpha Level of Significance**
Individual Stops	0.102452	4.35E-11	α= <1%
FT%	5.451494	0.000827	α= <1%
ORtg	0.175948	5.65E-16	α= <1%
DRtg	-0.30385	7.05E-30	α= <1%
Stop%	17.13219	4.76E-07	α= <1%
eFG%	21.91347	4.46E-09	α= <1%
AST	0.020838	6.69E-07	α= <1%

This regression documents the large impact that one data entry error had on my model as the adjusted R^2 value increased to 0.714458. Re-including these variables increased the model's accuracy by approximately 10%. Additionally, these extra variables are highly significant at an α= <1%, which means that these variables are highly influential in measuring player performance. I was hopeful that reincluding these variables would be significant to the model's formulation, but I did not expect this level of significance for each of the reincluded variables or the associated increase in adjusted R^2 value. Additionally, the F value has decreased to 7.7E-115 which highlights the strength of the variables working together within the model given this high level of joint significance. This iteration was a great step forward for properly measuring player performance, but the current model has 4 offensive variables, FT%, ORtg, eFG%, and AST, and 3 defensive variables, Individual Stops, DRtg, and Stop%. I intended for my model to have an increased defensive emphasis, but this current formulation is too heavily weighted toward the defensive aspect of the game. Because of this desire for increased offensive variables within the model, I attempted to include Total Rebounds, represented as TRB, and reintroduce turnovers per 100 possessions, shown as TO/100 Poss, due to the important theoretical relevance of these

statistics to help players exert an impact on the game. These inclusions change the PRL as shown in Equation (26).

$$WAR = \beta 0 + \beta 1 * \text{Individual Stops} + \beta 2 * \text{ORtg} + \beta 3 * \text{Stop\%} + \beta 4 * \text{eFG\%} + \beta 5 * \text{AST} + \beta 6 * \text{DRtg} + \beta 7 * \text{FT\%} + \beta 8 * \text{TRB} + \beta 9 * \text{TO/ 100 Poss} \quad (26)$$

The results of these inclusions are shown in Table 10.

Table 10: Model Iteration 7 with 10 years of data

10 years of data	$R^2 = 0.714849$	N=440	F value=3.4E-113
Variables	**Coefficient**	**P-value**	**Alpha Level of Significance**
Individual Stops	0.103441	1E-08	α= <1%
FT%	6.6068797	0.00479	α= <1%
ORtg	0.152039	6.16E-09	α= <1%
DRtg	-0.3042	1.24E-29	α= <1%
Stop%	15.92705	7.59E-06	α= <1%
eFG%	23.83721	1.71E-09	α= <1%
AST	0.024799	4.27E-07	α= <1%
TRB	0.002103	0.548699	Insignificant
TO/100 Poss	-0.25742	0.110745	α= 12%

This newest inclusion actually increases the model's ability to explain the relationship between the input variables and the output to an adjusted R^2 =0.714849. In contrast, the F value decreased slightly from the last iteration. However, this decrease is largely irrelevant given the relatively small number as the value remains highly significant using an α= <1%. With the inclusion of the new variables, the rest of the variables did become slightly less significant than the previous iteration. Much like the F value, these changes were largely unimportant in the context of the model as these values remain quite separated from the alpha value of 1%. However, the TRB variable had a p-value of 0.548699, which suggests that the variable is insignificant in modeling player performance as the p-value is not close to any accepted values of alpha that would make the variable significant. Given the large amount of data analyzed in the model, I made the decision to remove the TRB variable from the model. The TO/100 Poss variable was significant

at an alpha level of 12%. This level of significance suggests an influence on the model, but this variable does not exert the same influence as the other variables that are more significant at lower levels of alpha. The new PRL is shown in Equation (27) below.

$$WAR = \beta 0 + \beta 1 * \text{Individual Stops} + \beta 2 * \text{ORtg} + \beta 3 * \text{Stop\%} + \beta 4 * \text{eFG\%} + \beta 5 * \text{AST} + \beta 6 * \text{DRtg} + \beta 7 * \text{FT\%} + \beta 8 * \text{TO/ 100 Poss} \quad (27)$$

Table 11 shows the results of the regression taken after the TRB variable is removed so that the PRL is modeled by Equation (27).

Table 11: Model Iteration 8 with 10 years of data

10 years of data	$R^2 = 0.715273$	N=440	F value = 3.4E-114
Variables	**Coefficient**	**P-value**	**Alpha Level of Significance**
Individual Stops	0.108401	1.52E-11	α= <1%
FT%	5.728274	0.000477	α= <1%
ORtg	0.155131	1.46E-09	α= <1%
DRtg	-0.30573	3.53E-30	α= <1%
Stop%	16.51958	1.35E-06	α= <1%
eFG%	23.8129	1.72E-09	α= <1%
AST	0.023892	3.23E-07	α= <1%
TO/100 Poss	-0.2325	0.035639	α= 4%

This iteration of the model has a slight increase in the adjusted R^2 value as the value increased to 0.715273. However, this change is very small which suggests that removing the TRB variable was the correct decision as the variable does not have a noticeable impact. This removal upholds the model's ability to explain the relationship between variables at a rate of approximately 71.5%. The F value also slightly decreases. This decrease means that the variables are more jointly significant than the previous version of the model which suggests that this combination of variables is more effective in explaining the changes in player performance. With the exception of the TO/100 Poss variable, all variables are significant at an alpha level of less than 1% which suggests that these statistics are highly determinant in describing what causes high levels of

player performance. In contrast, the TO/100 Poss variable is only significant at an alpha level of 4%, which suggests that this variable does have an important influence on player performance but not as influential as the other variables. Given this higher level of alpha, I considered removing the variable, but I felt that this variable has theoretical importance. Additionally, this variable is statistically significant, albeit less so than the other variables, which merits the inclusion of the variable given the size of the data sample that is analyzed. I was surprised that this variable became significant after ten years of data given its insignificance after six years. I believe this result supports the need for a larger sample size to smooth out any possibility for increased variability.

As mentioned previously in the paper on page 23, multicollinearity and heteroskedasticity are two key issues that could alter my model's ability to properly estimate the influences of the variables in the model. Recall from earlier in the paper, multicollinearity is one variable having a direct linear relationship with another variable in its formulation which causes a high level of correlation between the variables. This issue is an important consideration for my model as multicollinearity could cause the variables in the model to inaccurately represent the true impact in determining which variables were essential to properly evaluate player performance to highlight top performers. There are two methods of testing for multicollinearity, which are a correlation matrix and variance inflation factors, or VIFs. A correlation matrix highlights the relationship between each pair of variables within the model. Commonly, a value of 0.7 or greater is considered evidence of multicollinearity. The correlation matrix used to test for multicollinearity is shown in Table 12.

Table 12: Correlation Matrix

	Indiv Stops	*ORtg*	*DRtg*	*Stop%*	*eFG%*	*AST*	*FT%*	*TO/ 100 poss*
Indiv Stops	1							
ORtg	0.14987	1						
DRtg	-0.3790	-0.16758	1					
Stop%	0.01782	0.139484	-0.0433	1				
eFG%	0.19097	0.72198	-0.2071	0.0103	1			
AST	0.00607	0.08956	0.03680	0.3858	-0.12407	1		
FT%	-0.2546	0.31632	0.10647	0.2658	0.066064	0.207778	1	
TO/ 100 poss	0.1729	-0.42667	-0.0448	-0.0174	-0.17766	0.27955	-0.1315	1

In Table 12, the only pair of variables with a correlation value of 0.7 or greater is eFG% and ORtg. Fundamentally, this finding makes sense as effective and efficient shooting leads to a higher offensive rating. However, I want to eliminate severe multicollinearity within this model, so this development is somewhat concerning. To fully determine the extent of the multicollinearity, I analyzed the VIFs for each variable to determine the true presence of multicollinearity within the model. This analysis is achieved by regressing the independent variables against one another while rotating which independent variable is considered as the output of the model. The R^2 value for each of these regressions is used to calculate VIF by the following equation:

$$VIF = 1/ (1- R^2) \quad (28)$$

In this context, a VIF value of greater than 5 exhibits the presence of severe multicollinearity. Lower VIF values suggest that multicollinearity is not present within the model. These VIF findings is shown in Table 13.

Table 13: VIFs

Variable	R^2 value when considered the response variable	VIF value
Individual Stops	0.275442	1.380151
ORtg	0.734388	3.764892
DRtg	0.169838	1.204584
Stop%	0.205203	1.258183
eFG%	0.625029	2.666873
AST	0.363051	1.569985
FT%	0.287067	1.402656
TO/ 100 Poss	0.423917	1.73586

As shown in Table 13, none of the VIF values are greater than 5 which suggests that multicollinearity is not present in the model. Despite the high correlation value between eFG% and ORtg as shown in the correlation matrix, I believe that this relationship is definitional and does not require further attention. Because of this result, I conclude that multicollinearity is not an issue within the model.

Another issue that prevents econometric models from being accurate is heteroskedasticity. Recall from earlier in the paper, heteroskedasticity occurs as the variance of the error terms varies with the independent variables which violates the classical assumption that the population errors are constant. Heteroskedasticity could affect my model by causing the variables to inaccurately measure the key statistics that I have outlined. To determine the effects of heteroskedasticity, the two tests typically used are the Park Test and the White Test. To use the Park Test, the residual values of the regression are found by the Equation (29) below.

$$\text{Residual value} = e_i = \text{expected output from the data point}_i - \text{actual output} \quad (29)$$

After finding these residual values, these values are squared and then the natural log of the squared values is taken, represented as $\ln(e_i^2)$. Additionally, the natural log is taken of each of the independent variables. Then, the natural log of the residual values is regressed against each

of the natural log of the independent variables individually to determine the effects of heteroskedasticity. The results of the Park Test are shown in Table 14 below.

Table 14: Park Test Results

LN (variable)	**R^2**	**P-value**
LN(Individual Stops)	0.001866	0.36596
LN(ORtg)	0.000304	0.715345
LN(DRtg)	0.001758	0.380299
LN(Stop%)	0.002546	0.290958
LN(eFG%)	0.001114	0.484939
LN(AST)	0.000113	0.824287
LN(FT%)	2.22E-09	0.999214
LN(TO/100 Poss)	0.001109	0.485877

Table 14 shows the regressions between the natural log of the squared residuals do not have a strong relationship with any of the natural logs of the independent variables. The extremely low R^2 values show that these regressions have a poor fit. Additionally, the high p-values are far from any accepted levels of alpha that leads to the conclusion that heteroskedasticity is not present which upholds the assumption that the population errors are constant within the model. To ensure that heteroskedasticity is not present in the model, I also applied the White Test to my model. The White Test regresses the squared errors, e_i^2, against the independent variables of the model, the products of the independent variables, and the squares of the independent variable terms. Because I have so many independent variables, I utilize a truncated version of the White Test that includes the independent variables with the highest possibility for heteroskedasticity, albeit none of the variables showed evidence of heteroskedasticity, from the results of the Park Test. The variables included in this truncated White Test are Individual Stops, DRtg, and Stop%. This application of the White Test is shown in Equation (30) below.

$$e^2 = \beta_0 + \beta_1 * \text{Indiv Stops} + \beta_2 * \text{DRtg} + \beta 3 * \text{Stop\%} + \beta 4 * (\text{Indiv Stops} * \text{DRtg}) + \beta 5 * (\text{Indiv Stops} * \text{Stop\%}) + \beta 6 * (\text{DRtg} * \text{Stop\%}) + \beta 7 * (\text{Indiv Stops})^2 + \beta 8 * (\text{DRtg})^2 * \beta 9 (\text{Stop\%}^2) \quad (30)$$

After running this regression with n =440 observations, this model had an R^2 value of 0.070919. In the White Test, the decision statistic, χ^2, is calculated by n* R^2. In this regression, the value of n R^2= (440) * (0.070919) = 31.20436. The highest published value in the Chi-square distribution table is at an α= 1% and 100 degrees of freedom has an associated Chi-square value of 135.81, so my data has n = 440 observations with 439, which is n -1, degrees of freedom which suggests that the Chi-square value would be significantly higher than 135.81. Because my calculated Chi-square value is 31.20436 which is significantly less than 135.81, I conclude that heteroskedasticity is not present within the model by the White Test. This discovery further shows that heteroskedasticity does not have an effect on my model.

After analyzing multicollinearity and heteroskedasticity, the final PRL remains the same as shown in Equation (27). Substituting the coefficients from Table 10 into the final PRL, the completed model is presented in Equation (31) below.

WAR = -7.18901 + 0.108401* Individual Stops + 5.728274 * FT% + 0.155131 * ORtg + (-0.30573) * DRtg + 16.51958 * Stop% + 23.8129 * eFG% + 0.023892 * AST + (-0.2325) *TO/ 100 Poss (31)

Equation (31) displays the role that each variable in the model has in more accurately modeling player performance within the ASUN Conference to better select the all-conference teams.

After constructing the model from running regressions to analyze ten years of ASUN conference data, I utilized Microsoft Excel to determine the ten players each year who had the largest WAR values as calculated by my model. After determining the top performing players for each year, I compared these findings to the players that were originally selected by the coaches and media of the ASUN conference to the all-conference teams for their respective seasons.

Table 15 below exhibits the number of new players that were selected to the all-conference teams by the model but were not originally selected by the ASUN selection committee.

Table 15: New Players Selected to the All-Conference Teams

Season	# of new players on the all-conference team
2011-12	6*
2012-13	6
2013-14	4
2014-15	6
2015-16	6
2016-17	5
2017-18	2
2018-19	5
2019-20	4
2020-21	6
2021-22	4

It is important to note that 2011-12 was unique as 11 players selected to the 1st- and 2nd all-conference teams as two players tied for the last spot on the 1st team. As exhibited in Table 15, 50 players were not initially selected to be on their respective season's ASUN all-conference team but were worthy of being selected as one of the top ten players in the conference. Alternatively, Table 15 also suggests that 51 players were correctly selected which suggest a 50.5%, 51/101, success rate in picking the correct players to the team each year. This rate suggests that the current method is selecting the deserving players slightly more than half of the time. Consequently, this figure needs to increase for these selection committees to be respected enough so that they can continue having the opportunity to select the all-conference teams.

Additionally, some movement between the all-conference teams occurred as players shifted from the first team to the second team, second team to the first team, into the first or second team from not being selected, or not being chosen by the model despite being originally chosen for the all-conference teams which is shown in Table 15. In addition to changes in the selected all-conference teams, the conference player of the year also changed through the model's analysis. The conference player of the year was selected by choosing the player with the highest WAR value through the model's formulation and comparing this player to the selection committee's conference player of the year. These comparisons are shown in Table 16 below.

Table 16: Movement within the All-Conference Teams

Season	No change in selection	From 1st to 2nd	From 2nd to 1st	From unselected to 1st / 2nd	From 1st / 2nd to unselected	Did the Player of the Year change?
2011-12	4	1	0	6	6	Yes
2012-13	2	0	2	6	6	Yes
2013-14	2	2	2	4	4	Yes
2014-15	3	1	0	6	6	No
2015-16	3	1	0	6	6	Yes
2016-17	1	2	2	5	5	Yes
2017-18	6	1	1	2	2	Yes
2018-19	3	2	0	5	5	Yes
2019-20	4	1	1	4	4	No
2020-21	3	0	1	6	6	Yes

These results suggest that my analysis has been effective in highlighting biases toward the offensive aspect of the game of basketball. In seven out of ten years, there is significant movement of players into and out of the model which suggests the need for a reevaluation of the methods for selecting the players on the all-conference teams. However, the movement between the first and second teams was limited within the model. This observation suggests that better analysis needs to occur at the margin of the selections when determining whether a player should be selected for the second team or not selected for either team.

After completing the construction of the model and considering movements within the all-conference teams historically, the model was applied to this year's data. Using the same process as earlier, I determined which players should been selected to the all-conference teams and compared these results to the all-conference teams that were selected by the league. I have included these results in Table 17.

Table 17: 2021-2022 Results

	Selected Player	School	Model Selection	School
1st team	Darius McGhee	Liberty	Kyle Rode	Liberty
	Darian Adams	Jacksonville State	Kevin Samuel	Florida Gulf Coast
	Tavian Dunn-Martin	Florida Gulf Coast	Darius McGhee	Liberty
	Kevion Nolan	Jacksonville	Tyreese Davis	Jacksonville
	Ahsan Asadullah	Lipscomb	Juston Betz	Bellarmine
2nd team	Dylan Penn	Bellarmine	Shiloh Robinson	Liberty
	CJ Fleming	Bellarmine	Kevion Nolan	Jacksonville
	Kevin Samuel	Florida Gulf Coast	Damaree King	Jacksonville State
	Kyle Rode	Liberty	Tavian Dunn-Martin	Florida Gulf Coast
	Chase Johnston	Stetson	Darian Adams	Jacksonville State

The predictions from my model for this year suggest that the ASUN committee selected 6 of the top 10 players correctly for the all-conference teams this year. The four players who were not initially selected by the committee are highlighted in Table 17. The model was correct in 60% of its selections this year which is an improvement from the 50.5% success rate for the 10 years of data that I previously analyzed. Despite a small sample size, I think this year suggests an improvement in the selection committee's ability to select the true ten best performing players. By including this year's selections in considering the selection committee's recent history, their success rate increases to 51.35%, 57/111, which represents a small improvement in the committee's ability to accurately select the all-conference teams. The movement in player's

selection to the all-conference teams between the first and second teams is reflective of the historic trends shown in Table 15. In this year's selections, Kyle Rode moved from the second team to the first team as Darian Adams and Tavian Dunn-Martin moved from the second team to the first team. Additionally, Ahsan Asadullah, Dylan Penn, CJ Fleming, and Chase Johnston moved out of the first and second team selections as their spots were taken by newcomers to the model in Tyreese Davis, Juston Betz, Shiloh Robinson, and Damaree King. Additionally, Kyle Rode replaced his Liberty teammate, Darius McGhee, as Conference Player of the Year with the highest WAR value as calculated by the model.

Reflection:

The emphasis of my model from the very beginning was on the defensive side of the game, and I think this model achieves that goal. I wanted to move away from considering only the offensive side of the game, but I believe that this model could be an overcorrection towards measuring for more defensive impact than is present within the game. I wanted my model to consider the most efficient shooters and avoid rewarding volume shooters for accumulating many points simply through shooting many of their team's attempts or piling up points at the free throw line exclusively. However, I think my model may actually bias against volume shooters as the model does not account for points scored in any fashion. Originally, I wanted my model to avoid biasing my model toward the players on the top teams in the conference. I believe that my methodology in this paper attempts to avoid this pitfall, but there are several instances where one team has multiple players on the all-conference teams in a particular season. Despite this result, I contend that these teams were composed of the top players within the conference as they helped their respective teams achieve a high level of success that season. This idea suggests that the top players make the top teams and not the other way around, which discounts the theory about all-

conference teams being biased toward top teams. I was surprised that the conference player of the year changed in eight out of the ten years as shown in Table 16. However, I believe this speaks to the intent behind the model as the model aims to have less of a bias toward the offensive parts of the game. Marginal changes in the methods of analysis can decide whether a player is conference player of the year or only selected to the first or second team. This observation is important for a player who scores twenty points a night with lackluster defensive performance as a consideration of all aspects in the game of basketball is necessary to most accurately select the all-conference team.

I was pleasantly surprised by the R^2 value of my model given the amount of raw data that was being considered in my analysis. These results helped me to determine what I believe are the most essential variables in determining player performance. However, I was surprised by the lack of multicollinearity and heteroskedasticity in the model given the interconnectedness and flowing nature of the game of basketball. The sport is not heavily segmented like baseball or football, so I expected the variables to potentially dilute the effects of the other variables. I believe that the lack of these econometric issues displays the thoughtfulness put into selecting these variables to prevent these issues from occurring. Additionally, the lack of these issues suggests that the model is accurate in determining who the top players are within the ASUN conference to prevent the incorrect selections for the all-conference teams in the future.

Overall, I think my model presents a unique approach to evaluating player performance to better select the all-conference teams in the ASUN conference. Econometric models have been utilized previously in modeling collegiate basketball, but the inclusion of the Individual Stops and Stop% statistics allows for a more accurate analysis of player performance to truly determine which players have the greatest positive impact on their team's success. The model presented

within this paper attempts to emphasize the other aspects of the game of basketball without discounting a player's ability to score points at a high level. I believe the ideas presented in this paper could be utilized by other conferences in selecting their all-conference teams, even conferences that select a third all-conference team, all-defensive teams, all-freshmen teams or be used by college programs to evaluate players in the transfer portal to determine who would provide the most additional value to their team outside of the ability to score many points through volume shooting. Through my analysis, I realized how the coaches and media of the selection committee attempts to spread out these selections throughout the teams in the conference. However, this paper argues the need for the stoppage of this practice so that these all-conference teams truly represent the players who performed at the highest level for a particular season.

References

"ACC Men's Basketball Awards Announced." *ACC Men's Basketball Awards Announced - Atlantic Coast Conference*, ACC Network, 8 Mar. 2021, https://theacc.com/news/2021/3/7/acc-announces-mens-basketball-awards.aspx.

Annis, D. H. (2006) Optimal end-game strategy in basketball. Journal of Quantitative Analysis in Sport, 1030, 1-9

Atlantic Sun Conference. "ASUN Conference Men's Basketball Record Book." *Atlantic Sun Conference*, NCAA, 2020, asunsports.org/sports/mbkb/2019-20/files/ASUN_Mens_Basketball_Record_Book.pdf.

Bartholomew, James T., and Collier, David A. "The Role of Contested and Uncontested Passes in Evaluating Defensive Basketball Efficiency." *Journal of Service Science (Online)*, vol. 4, no. 2, 2011, pp. 33. *ProQuest*, https://libproxy.bellarmine.edu/login?url=https://www-proquest-com.libproxy.bellarmine.edu/docview/1418717087?accountid=6741.

Baumer, Benjamin S., Jensen, Shane T., Matthews, Gregory J. "OpenWAR: An Open Source System for Evaluating Overall Player Performance in Major League Baseball." *Journal of Quantitative Analysis in Sports*, vol. 11, no. 2, June 2015, pp. 69–84. *EBSCOhost*, search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=103105201&site=ehost-live&scope=site.

Berri, D. (1999) Who is 'most valuable'? Measuring the player's production of wins in the national basketball association, Managerial and Decision Economics, 20, 411–27.

Cochran, J. J. and Blackstock, R (2009): "Pythagoras and the national hockey league," *Journal of Quantitative Analysis in Sports*, Vol. 5: Iss. 2, Article 11

Engler, Mitchell L. “The Untaxed King of South Beach: LeBron James and the NBA Salary Cap.” *San Diego Law Review*, vol. 48, no. 2, Spring 2011, pp. 601–621. *EBSCOhost*, search-ebscohost-com.libproxy.bellarmine.edu/login.aspx?direct=true&db=a9h&AN=63485299&site=ehost-live&scope=site.

Ferreira, A. P., Volossovitch, A., & Sampaio, J. (2014) Towards the game critical moments in basketball: a grounded theory approach. International Journal of Performance Analysis in Sport, 14, 428-444.

Franks, Alexander, D’Amour, Alexander, Cervone, Daniel, Bornn, Luke. *Meta-Analytics: Tools for Understanding the Statistical Properties of Sports Metrics.* CFornell University Library, arXiv.org, Ithaca, 2016. *ProQuest*, https://libproxy.bellarmine.edu/login?url=https://www.proquest.com/working-papers/meta-analytics-tools-understanding-statistical/docview/2080371786/se-2?accountid=6741.

Gómez, Miguel Ángel, Lorenzo, Alberto, Jiménez, Sergio, Navarro, Rafael M., Sampaio, Jaime. “Examining Choking in Basketball: Effects of Game Outcome and Situational Variables during Last 5 Minutes and Overtimes.” *Perceptual & Motor Skills*, vol. 120, no. 1, Feb. 2015, pp. 111–124. *EBSCOhost*, search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=101158767&site=ehost-live&scope=site.

Henson, R. K. (2002). The logic and interpretation of structure coefficients in multivariate general linear model analyses. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA

Hoffmann, M.., McEwan D., Baumeister, R., Barnes, J, Guerrero, M. "Home Team (Dis)Advantage Patterns in the National Hockey League: Changes Through Increased Emphasis on Individual Performance With the 3-on-3 Overtime Rule." *Perceptual & Motor Skills*, vol. 128, no. 1, Feb. 2021, pp. 424–438. *EBSCOhost*, doi:10.1177/0031512520966138.

James, B. (1980): Baseball Abstract, self-published.

Kilanowski, Humbert. "SABR." *Society for American Baseball Research*, Sabr /Wp-Content/Uploads/2020/02/sabr_logo.Png, 16 June 2020, https://sabr.org/journal/article/cwar-modifying-wins-above-replacement-with-the-cape-cod-baseball-league/.

Major League Baseball. "What Is a Wins Above Replacement (WAR)?: Glossary." *Major League Baseball*, m.mlb.com/glossary/advanced-stats/wins-above-replacement.

Manner, Hans. "Modeling and Forecasting the Outcomes of NBA Basketball Games." *Journal of Quantitative Analysis in Sports*, vol. 12, no. 1, 2016, pp. 31–41., doi:10.1515/jqas-2015-0088.

Mason, Daniel S., and William M. Foster. "Putting Moneyball on Ice?" *International Journal of Sport Finance*, vol. 2, no. 4, Nov. 2007, pp. 206–213. *EBSCOhost*, search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=27989828&login.asp&site=ehost-live&scope=site.

Molodchik, Mariia, Paklina, Sofiia, Parshakov, Petr. "Peer Effects on Individual Performance in a Team Sport." Journal of Sports Economics, vol. 22, no. 5, June 2021, pp. 571–586, doi:10.1177/15270025211000389.

Myers, Daniel. “About Box Plus/Minus (BPM).” *Basketball Reference*, Feb. 2020, www.basketball-reference.com/about/bpm2.html.

Oliver, Dean. *Basketball on Paper: Rules and Tools for Performance Analysis*. Vol. 1st ed, Potomac Books, 2004. *EBSCOhost*, search-ebscohost-com.libproxy.bellarmine.edu/login.aspx?direct=true&db=nlebk&AN=388578&site=ehost-live&scope=site.

Pérez, Levi. “Will We Lose If We Lose You? Players’ Absence, Teams’ Performance and the Overlapping of Competitions.” *Journal of Sports Economics*, vol. 22, no. 6, Aug. 2021, pp. 722–734, doi:10.1177/15270025211008499.

Shea, Stephen M., and Baker, Christopher E. “Calculating Wins over Replacement Player (WORP) for NHL Goaltenders.” *Journal of Quantitative Analysis in Sports*, vol. 8, no. 1, Mar. 2012, p. 0. *EBSCOhost*, search-ebscohost-com.libproxy.bellarmine.edu/login.aspx?direct=true&db=s3h&AN=84343589&site=ehost-live&scope=site.

Schuckers, Michael E. and Curro, James. “Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events.” (2013).

Slowinski, Piper. “Replacement Level.” *Replacement Level | Sabermetrics Library*, Fangraphs, 26 Feb. 2010, library.fangraphs.com/misc/war/replacement-level/.

Studenmund, A. H. *Using Econometrics: A Practical Guide*. Pearson, 2017.

Weaving, Dan, Jones, Ben, Ireton, Matt, Whitehead, Sarah, Till, Kevin, Beggs, Clive. "Overcoming the Problem of Multicollinearity in Sports Performance Data: A Novel Application of Partial Least Squares Correlation Analysis." *PLoS One*, vol. 14, no. 2,

2019. *ProQuest*, https://libproxy.bellarmine.edu/login?url=https://www.proquest.com/scholarly-journals/overcoming-problem-multicollinearity-sports/docview/2180994330/se-2?accountid=6741, doi:http://dx.doi.org/10.1371/journal.pone.0211776.

Woolner, K. (2002): "Understanding and measuring replacement level," in Joe Sheehan ed., Baseball Prospectus 2002, Brassey's Inc: Dulles, VA, 55–66. *Measurement*

Yeatts, Paul E., Barton, Mitch, Henson, Robin K., Martin, Scott B. "The Use of Structure Coefficients to Address Multicollinearity in Sport and Exercise Science." *in Physical Education & Exercise Science*, vol. 21, no. 2, Apr. 2017, pp. 83–91. *EBSCOhost*, search-ebscohost-com.libproxy.bellarmine.edu/login.aspx?direct=true&db=s3h&AN=122101221&login.asp&site=ehost-live&scope=site.