

Analyzing NBA Player Performance using Regularized Adjusted Plus/Minus

Mohamed Hassan El-Serafi

Table of Contents

Abstract	1
Introduction	2
Literary Review	3
Advanced Statistics over Traditional Statistics	3
Other Advanced Metrics Applied	4
Variations of Plus/Minus	4
Creation of Regularized Adjusted Plus/Minus and Alternatives to RAPM	5
Critiques of Plus/Minus Usage in Player Evaluation	7
Methodology	7
Results	11
Ridge Regression	11
Lasso	15
Elastic Net	19
Bayesian Linear Regression	23
Discussion	28
Conclusion	30
References	31

Abstract

This analysis attempts to expand on previous studies that transform the Plus/Minus statistic into a reliable stat to evaluate NBA players. Using play-by-play data from the last four completed NBA regular seasons (2020-21 to 2023-24), multiple regularization models (Ridge, Lasso and Elastic Net) and a Bayesian Linear Regression model were applied to compare the 30 players with the highest plus/minus values which were derived from their respective coefficients. The models were also applied to data that was filtered to include only players who played the median amount of minutes. Among the top 30 players with the highest coefficients,

Joel Embiid and Nikola Jokic ranked 1st or 2nd in each model. There were no significant variations in player coefficient values between the aggregated and filtered datasets for the top 30 players. However, Lasso and Elastic Net produced higher player coefficient values than Ridge, while the Bayesian models produced the lowest player coefficient values, with notable changes in rank among players with the top 30 coefficient values. Model performance were assessed using Root Mean Squared Error (RMSE) and Adjusted R-squared. There was not a considerable difference in model quality for either metric, with RMSE values ranging from 69.868 to 70.013 and Adjusted R-squared values ranging from 0.003 to 0.008. Because of the parameters used and its computationally-intensive process, both Bayesian models did not generate Adjusted R-squared results, but did have the highest RMSE among all the models.

Introduction

Organizations in the National Basketball Association have attempted to evaluate player performance using various methods. Scouts attend basketball games to assess a player's strengths and weaknesses. Video coordinators utilize film from previous games to analyze how their own respective team or opposing team's players performed, building reports intended to highlight areas of strengths and weaknesses, and strategize how best to help improve the performance of their own players or approach playing against an opposing team's players. Traditional basketball statistics such as points, rebounds, assists, turnovers, field goal and 3-point percentages are commonly used as metrics to evaluate a player's on-court productivity. While these evaluation methods are still applied, advanced statistics has become increasingly utilized to assess a player's performance. This analysis focuses on the Regularized Adjusted Plus/Minus statistic. It is an advanced stat that is a transformed version of the Adjusted Plus/Minus, which is derived from raw plus/minus stats of players. Raw plus/minus stats were first created by Andy Hill in 1976, who was the coach of the Men's College Basketball team at Santa Monica Community College at the time (DeCourcy 2024). Terming the stat as the "Team Contribution Index", Hill's premise for the statistic was this: how does the team perform when Player A is in the game? How many points were scored for and against a team while Player A was active in the game? The definition for the calculation is simple: a specific player's plus/minus is the difference between team points scored and team points allowed while the player is on the court

$$(\text{TeamPointsScoredWhileThatPlayerisOnTheCourt}) - (\text{TeamPointsAllowedWhileThatPlayerisOnTheCourt})$$

("NBA Plus-Minus & Impact Metrics in Basketball Explained" 2017). As Hill states, "...this actually encourages team play. It gets people to think how the team is doing when I'm in the game, which is not natural for some people(DeCourcy 2024)." Plus/Minus provides player evaluators the ability to view player performances through the prism of how they impact a team's performance while they were in the game. Additionally, the statistic provides insight into how a team performs without the player on the court. For example, if Team A outscore their opponents by 15 points during Player X's 25 minutes of play, but is outscored by 5 points during Player X's 23 minutes off the court, Player X would have a plus/minus of +10 for that

game. As I will explain later, there are flaws with using raw plus/minus stats, with different variations of the statistic that have been created to mitigate its statistical shortfalls when measuring a player's impact on a team's performance.

Literary Review

Advanced Statistics over Traditional Statistics

The preference for applying advanced statistics to player evaluation comes from the belief that traditional statistics does not adequately capture a player's productivity on the basketball court. An example of the deviation from traditional to advanced statistics is the reliance on Effective Field Goal Percentage (eFG%) and True Shooting Percentage (TS%) over traditional Field Goal (FG%) and 3-point field goal (3PT-FG%) percentages to evaluate a player's shooting efficiency. Field Goal percentage is determined by a simple equation: the number of made field goals divided by the number of total field goal attempts. Similarly, 3-point and free throw percentages are calculated by the number of made baskets divided by the total number of attempts, respectively. Prior to 1979, FG% made complete sense, as only two point field goal attempts existed (Jacobs 2017). The idea of high probability shots referred to regions on the court that would lead to the highest probability of scoring two points. However, with the introduction of the three point line in 1979 and its subsequent increase in overall player shot attempts, the use of field goal percentage as the main barometer for assessing a player's shooting ability has become less dependable. To incorporate the prominence of 3-point stats, eFG% was created, using the formula

$$(FieldGoalsMade + (0.5 * 3PointFGM) / FieldGoalAttempts)$$

. The purpose is to give more weight towards 3-point field goals made, weighing the number of field goals made by 1.5 if the field goal made is a three-point attempt. Additionally, eFG% captures scoring efficiency per field goal attempt in an unbiased manner in comparison to FG%. For instance, if Player A and Player B each attempt 1,000 field goals and make 500, but Player B had 40 three-point field goals made, Player B's effective field goal percentage would be 52% compared to Player A's 50%, with Player B having scored more points (1,030) than Player A (1,000). True Shooting Percentage (TS%) attempts to take this idea a step further by introducing a "points per shooting possession" model. TS% attempts to capture the effect of free throw attempts combined with total field goal attempts and points scored, using the formula

$$Points / 2 * (FGA + (.44 * FTA))$$

. By incorporating free throw attempts and total field goal attempts (2-point and 3-point attempts) with points scored, the statistic captures all aspects of a player's opportunities to score and provide a better evaluation metric to consider than traditional 2PFG%, 3PFG%, and FT% stats.

Other Advanced Metrics Applied

Player Efficiency Rating (PER) and Net Rating are two other common advanced metrics used to measure a player's effectiveness impacting a basketball game. According to basketball-reference.com, PER was created by John Hollinger, and is a statistical rating of a player's per-minute productivity, adding up all of a player's positive accomplishments all a player's positive accomplishments, subtracting their negative accomplishments, and returning a per-minute rating of a player's performance ("Calculating PER" n.d.). Net rating calculates the offensive rating

$$(100 * (Points)/(Possession))$$

minus the defensive rating

$$(100 * (OpponentPoints)/(Possession))$$

, indicating how much better or worse the team performs with the player on the court (Fijal n.d.). These ratings are typically calculated on a per-possession basis.

Variations of Plus/Minus

The Plus/Minus statistic has become widely used by player evaluators to help inform their decision-making on roster construction and game strategy. It indicates how many points a team gained or lost relative to the opponent when a player was on the floor (Deckard 2024). For example, if Player A comes in the game and his team is up by 10, but they're tied when he exits the game, his plus/minus for that stretch will be -10. While this stat may give insight on how a player positively or negatively impacts their team's performance, there are other variables that taken in its raw form does not account for, namely the fact that there are four other offensive players and five defensive players on the court that could have more of an impact on the game, whether the player is playing at home or on the road, and whether the minutes a player plays are important based on the time and score of the game. This raises some questions: Does a player with an unexpectedly high Plus/Minus benefit from circumstances, or is he doing something that creates value that we might not be seeing? Is a player with an unexpectedly low Plus/Minus not as good as we thought or is he just in an unfortunate situation (Antle n.d.)?

There have been variations of Plus/Minus created to provide better player performance measurements. Box Plus/Minus (BPM), a metric developed by Daniel Myers, estimates a player's contribution to the team when that player is on the court ("About Box Plus/Minus (BPM)" n.d.). It is based only on the information in the traditional basketball box score—no play-by-play data or non-traditional box score data (like dunks or deflections) are included. The metric uses a player's box score information, position, and the team's overall performance to estimate the player's contribution in points above league average per 100 possessions played. The calculation includes a player box score stats per 100 possessions, team-adjusted efficiency

per 100 possessions, and estimated player positions (Go 2025). BPM evaluates a player’s performance in relation to the league average, set at 0.0. This means a player with a BPM of +5.0 would make his team 5 points per 100 possessions better than a league-average player would (admin 2024). While BPM provides an adequate method for measuring a player’s overall impact using traditional statistics, there are areas that are unaccounted for. As previously noted, BPM does not include play-by-play data, failing to capture the other nine players on the court and score at the time the player entered and left the game. Additionally, BPM can be influenced by “statistical noise”, as it uses team performance as a baseline but applies it to individual players. External factors, such as “garbage time” in which the margin of the game has been already decided, can also skew results, while the impact of teammates makes it unclear whether a player’s plus-minus reflects their contribution or that of their team (Go 2025).

To address the statistical and analytical shortcomings of the various Plus/Minus calculations, Dan Rosenbaum, Jeff Sagarin and Wayne Winston developed the Adjusted Plus/Minus (APM). The purpose of the APM stat reflects the impact of a player on his team’s scoring margin, factoring the other nine players on the court, the length of time a player is in the game, and the score at the start and end of the player’s time on the court (“Adjusted Plus-Minus (APM) Explained” 2017). In other words, it estimates player variables which produce the smallest difference between the expected margin and the actual margin in the matchups. It relies on weighted least squares, where you solve for the p -dimensional vector of player coefficients β using a modified version of the traditional least squares model (Yurko and Nguyen n.d.):

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{t=1}^T n_t (y_t - X_t \beta)^2$$

Although this metric comes close to an unbiased measure of a player’s effectiveness, there are conflicts within the regression model. There is high variance, which indicates that the metric is capturing noise in the data and is prone to overfitting. This can lead to drastic variations of a player’s adjusted plus/minus value from one season to the next. Additionally, APM is susceptible to multicollinearity due to how players can be used in pairs and/or trios within 5-player lineups.

Creation of Regularized Adjusted Plus/Minus and Alternatives to RAPM

Attempting to correct the statistical flaws in APM, Joseph Sill applied a regularization method to the model. The purpose for using a regularization method is its ability to handle multicollinearity and prevent overfitting by shrinking coefficients towards zero (Murel and Kavlakoglu 2023). The metric is called Regularized Adjusted Plus/Minus (RAPM). RAPM applies a Ridge regression model, reducing the standard errors from APM, correcting for multicollinearity and addressing overfitting in the training data. A Bayesian technique is used with Ridge regression as an enhancement in which the data is combined with prior knowledge

regarding reasonable ranges for the parameters to produce more accurate models (Sill 2010). Ridge regression provides a regularization term that is the sum of the squares of the model's coefficients. The formula utilized by Sill is

$$\sum_i p_i (t_i - \mathbf{w} \cdot \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|^2$$

where p_i is the number of possessions in the game snippet, t_i is the difference in offensive efficiency between the home and road teams for the game snippet, x is a vector of 0s, 1s, and -1s with the 1s representing the presence of the home players on the court and -1s representing the road players (there is also a constant 1 representing the ever-present home court advantage) (Sill 2010).

Sill used data of complete seasons from the 2006-07 and 2007-08 and part of the 2008-09 season as his training data to predict his test set, which are the last two months of the '08-'09 season (March and April). Additionally, a lambda value is used as a penalty to draw the coefficients close to zero, with a value of 2000 used for 1 year of data and 3000 for three years of data. Sill uses Root Mean-Squared Error (RMSE) of predicted vs. actual margin to measure the accuracy of the model, and cross-validation to determine the optimal minutes cutoff and weightings of past years for the standard APM linear regression technique. In Sill's analysis, he uses data from the 2006-07 season through February of the 2008-09 season as his training set, with March and April of the 2008-09 season as his test set. Separately, Sill also uses only the 2008-09 season as his dataset, with data through February 2009 as his training set. He explains that the Adjusted R-squared was lower and Root Mean Squared Error (RMSE) was higher when using only 1 season of data compared to using three seasons of data.

As an alternative to using RAPM, Damoulaki, Ntzoufras, and Pelechrinis developed a different method called weighted expected points (wEPTS) (Damoulaki, Ntzoufras, and Pelechrinis 2025). According to the authors, the new evaluation metric is based on the expected points per possession (EPTS) and its extended version, the weighted expected points (wEPTS). The latter measurement takes into consideration the proportion of possessions that each player participates in his team. It is derived as the points expected to be scored by the team over all games under the simplified scenario that all other players on court belong to the reference group with zero-lasso RAPM measures. In addition to using ridge and lasso regression models, the authors applied binary and multinomial logistic regression models. Their thinking is that the multinomial model is more suitable for modelling the points per possession and for producing player evaluation metrics since the obtained wEPTS evaluation index is found to be superior according to selected external validation criteria. The authors state that the metric partially solves the problem related to the overestimation of the performance of low-time players. The binary classification model was applied, predicting the outcome of scoring versus not scoring. Regularized logistic regression was implemented, achieving an accuracy of approximately 55%. However, this approach neglects the actual number of points scored on each possession. Despite this, the authors discovered a linear relationship between the ratings

generated from the ridge binomial model and ridge normal RAPMs. As a result, they converted normal RAPMs into logistic regression RAPMs, transforming the normal RAPM scores into a framework more appropriate for the binary outcome (scoring vs. not scoring), increasing interpretability of standard RAPM player performance metrics. The authors then applied a multinomial model to the number of points per possession by using three binomial models for the three scoring categories under consideration, accounting for how many points can be scored during a possession (1, 2 or 3+). This is used to develop their own metric, Expected Points Per Possession (EPTS), and weighted Expected Points Per Possession (wEPTS).

Critiques of Plus/Minus Usage in Player Evaluation

While Plus/Minus has undergone various modifications to address statistical flaws, there has been criticism with how credible the stat is when evaluating individual player performance. Jack Van Deventer argues that because there are four other players on the court that a player plays with, it is difficult to isolate that particular player's influence on the game (Van Deventer n.d.). Therefore, Van Deventer suggests that plus/minus should not be used to analyze individual players and instead should be applied to evaluate multi-player combinations (2, 3, 4, and 5-man lineups). This would produce better team insights while reducing variability. Second, the amount of minutes a player plays in a game can vary, with a player producing a high positive or negative plus/minus in a low amount of minutes/possessions. For instance, if Player X averages 5 minutes played per game and produces a plus/minus of +10 and Player Y averages 35 minutes and produces a plus/minus of +5, it may appear that Player Y has had less of an impact than Player X based on their respective plus/minus stat. However this is misleading, as the sample size of minutes for Player X is disproportionately small compared to Player Y, creating bias towards the player who played less minutes. Therefore, a larger sample size is recommended in order to adequately compare and evaluate player productivity.

Methodology

For this analysis, I applied the Regularized Adjusted Plus/Minus metric using play-by-play data for the past four completed NBA seasons: 2020-21 through 2023-24. Since a player's performance may vary from one season to the next, the data compiles all four seasons, creating one large dataset. The data was collected using an R package called hoopR. The process for extracting the data and implementing the models were derived from Ron Yurko and Quang Nguyen (Yurko and Nguyen n.d.). The input variables are the players, with the margin (defined as:

$$(homepoints - awaypoints)/npos * 100)$$

being the target variable. Each row represents a possession, with the player data transformed into matrix form. Each player is assigned a 0 or 1 based on whether the player was on the

court for the possession (1 for on the court, 0 for off). The formula was adapted from Yurko and Nguyen, and is used to estimate player coefficients:

$$\hat{\beta}^{ridge} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{t=1}^T (y_t - X_t \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

According to Yurko and Nguyen, the objective for the Ridge regression coefficients is effectively the combination of the loss (the traditional least squares objective) and newly included penalty term (the sum of the squared coefficient values). The Ridge regression coefficients are solved for while balancing these two terms simultaneously, with the amount penalization controlled by λ . We can consider λ to be a tuning parameter that controls the strength of the penalty term, and we will want to choose the λ based on out-of-sample performance (Yurko and Nguyen n.d.).

There are a couple of options available to account for players that do not play a lot of minutes. Damoulaki, Ntzoufras, and Pelechrinis discuss the concept of Low-Time Players (LTP), defined as players who participate for less than a threshold of minutes per game (Damoulaki, Ntzoufras, and Pelechrinis 2025). While the authors implemented a minimum of at least 200 minutes played, Sill applies a cross-validation method to determine the adequate minutes threshold for each player. The purpose of applying a type of threshold for minutes played is to account for the skewness in the amount of minutes players play in games. Sill highlights the minutes threshold for his model as being 1,200 minutes for a single-season analysis, and 1,300 minute threshold for a three-year analysis. There are a disproportionate amount of players who don't play a lot of minutes. This is reasonable, as there are only 5 players allowed to play on the court per team, with each team carrying 12-15 players at a time. The distribution of minutes played in this dataset illustrates this:

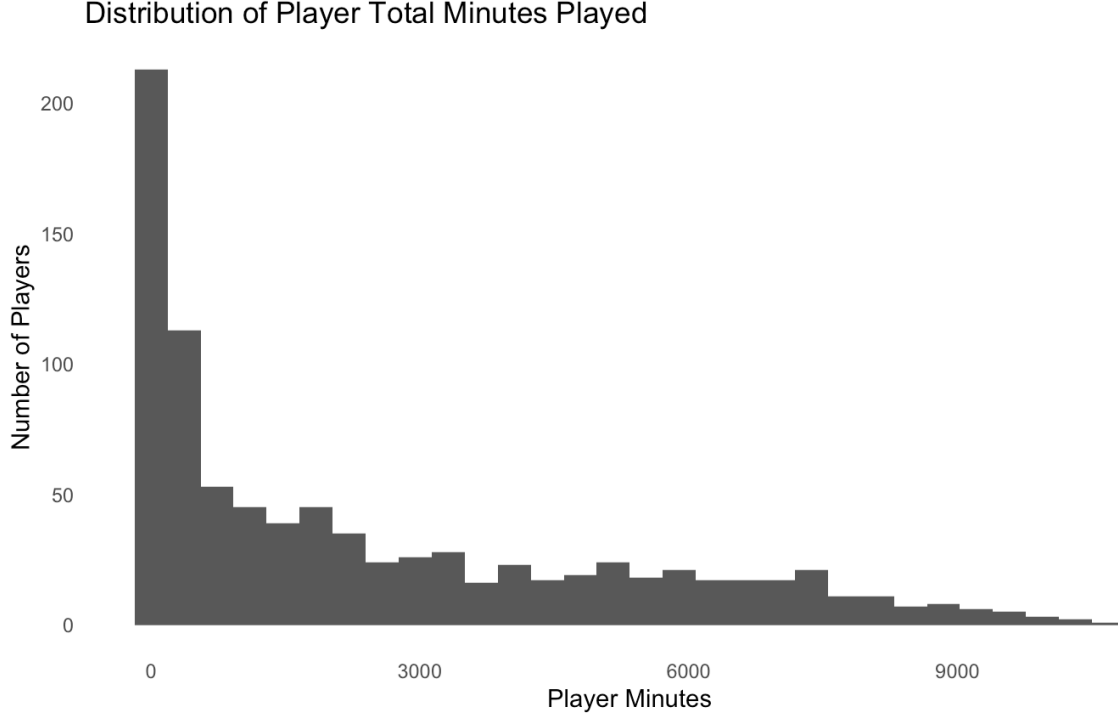


Figure 1: Distinct skewness in amount of minutes played.

Instead of the methods implemented by Sill and Damoulaki et al., I used the median amount of minutes in the dataset as the threshold for the amount of players to use. Because the data is skewed and contains outliers, using the median would be more appropriate to use than the mean for setting a minutes threshold (Bobbitt 2021). For each type of model, I used two separate dataset, one dataset containing the total number of players and the other containing players who played more minutes than the median.

The regression models adapted for this analysis are derived from methods applied by Sill and Damoulaki et al. Sill applied a Ridge regression model, while Damoulaki et al utilized a Lasso multinomial classification model. Unlike Damoulaki et al., however, the type of Lasso model used for this analysis is regression. Unlike Ridge, which uses the square of the magnitude of the coefficients, Lasso shrinks coefficients to zero, using the absolute value of the coefficient magnitude and performing a feature selection process that reduces the number of independent variables (“Ridge Regression Vs Lasso Regression” n.d.). The formula used is as follows:

$$\hat{\beta}^{lasso} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{t=1}^T (y_t - X_t \beta)^2 + \lambda \sum_{j=1}^p \beta_j$$

In addition to these two models, two other regression techniques are applied to the dataset:

Elastic Net and Bayesian Linear regression. The purpose for using Elastic Net regression is its use of both Ridge and Lasso techniques, combining their penalty terms and learning from their shortcomings to improve regularization of statistical methods (“Elastic Net” n.d.). It combines both L1 (Lasso) and L2 (Ridge) penalties to perform feature selection and manage multicollinearity, balancing coefficient shrinkage and sparsity:

$$\hat{\beta}^{EN} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{t=1}^T (y_t - X_t \beta)^2 + \lambda \left[\alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right]$$

This method helps reduce overfitting while retaining all features in the model, while providing a more stable and generalized model compared to using Lasso or Ridge alone (“Lasso Vs Ridge Vs Elastic Net | ML | GeeksforGeeks” n.d.). From previous analyses of RAPM, the authors have not applied an Elastic Net regression method on play-by-play data, and comparing its results to the other regularization models Ridge and Lasso could help give us a better understanding of each method’s strengths and limitations.

The purpose for using Bayesian regression is its ability to incorporate prior knowledge of the data. Since the data consists of prior multiple seasons of player data, Bayesian may be an appropriate model to utilize for this analysis. In fact, the regularization models utilize a Bayesian technique in which the data is combined with *a priori* beliefs regarding reasonable ranges for the parameters in order to produce more accurate models (Sill 2010). Priors on the parameters are used within the model as a lambda equivalent, where they can have a regularizing effect (Tim 2018). Chains and iterations are two other parameters used within the model. The purpose of using chain, which is Markov Chain Monte Carlo sampling method, is to sample the posterior distribution in groups, while the iterations iterate through each sample within the chain (“What’s in a Bayesian Model? | R” n.d.). While using Bayesian may not be suitable due to the large size of the dataset, it’s simply used to assess its results with the other models. The Bayesian formula used is displayed below:

$$\text{margin}_i \sim \mathcal{N} \left(\beta_0 + \sum_{j=1}^P \beta_j \cdot x_{ij}, \sigma^2 \right)$$

$$\beta_j \sim \mathcal{N}(0, 1) \quad \text{for } j = 1, \dots, P$$

$$\beta_0 \sim \mathcal{N}(0, 5)$$

To split the data for the regularization models, I applied a cross-validation method, using the default nfolds of 10. A larger nfolds value would not be appropriate for large size of the dataset (“Cross-Validation for Glmnet — Cv.glmnet” n.d.). The two primary metrics used to assess each model were Root Mean Squared Error (RMSE) and R-squared. These metrics were used by Sill to compare the different Ridge regression models he used when comparing the results

from the amount of seasons used in the dataset. Additionally, I compared each of the player's plus/minus results from each of the models, assessing their coefficient outputs, which are the plus/minus values.

Results

As previously mentioned, there were two separate models created for each regression method. The purpose is to compare how the models performed with all of the players in the dataset and separately including only the players who played more than the median, which was 1424.75. The histogram plots below are the distribution of the coefficient values for all of the players from the respective datasets. The bar plots display the top 30 players with the highest coefficient (plus/minus) values. Additionally, the table below breaks down the RMSE, R-squared, and Adjusted R-squared for each of the models. In total, there were 885 total players in the dataset, and 443 players who played at least the median amount of minutes. The margin of score is the target variable, with the Player ID numbers being the input variables. The input data was transformed into a matrix containing binary values, with 1 representing players on the court and 0 representing players not on the court. The lambdas used for each of the regularization models were determined using cross-validation that produced the best lambda value.

Ridge Regression

All NBA Players

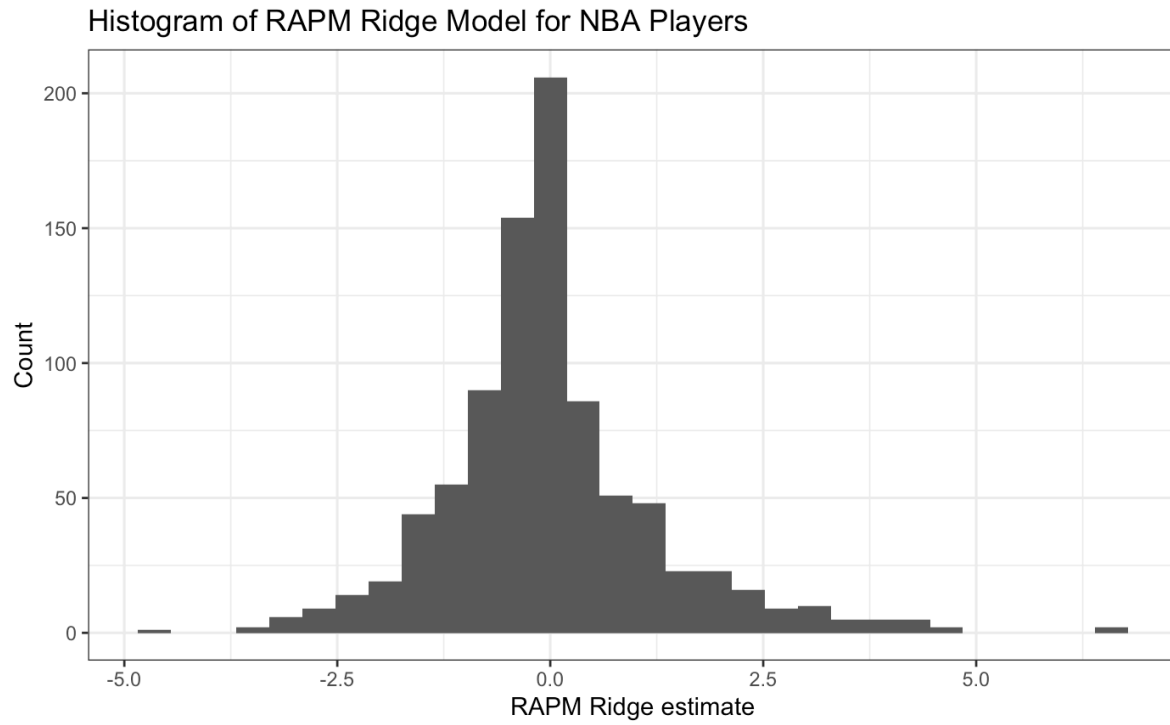


Figure 2: There is narrow spread of the coefficients, with its center near zero. This indicates low variability.

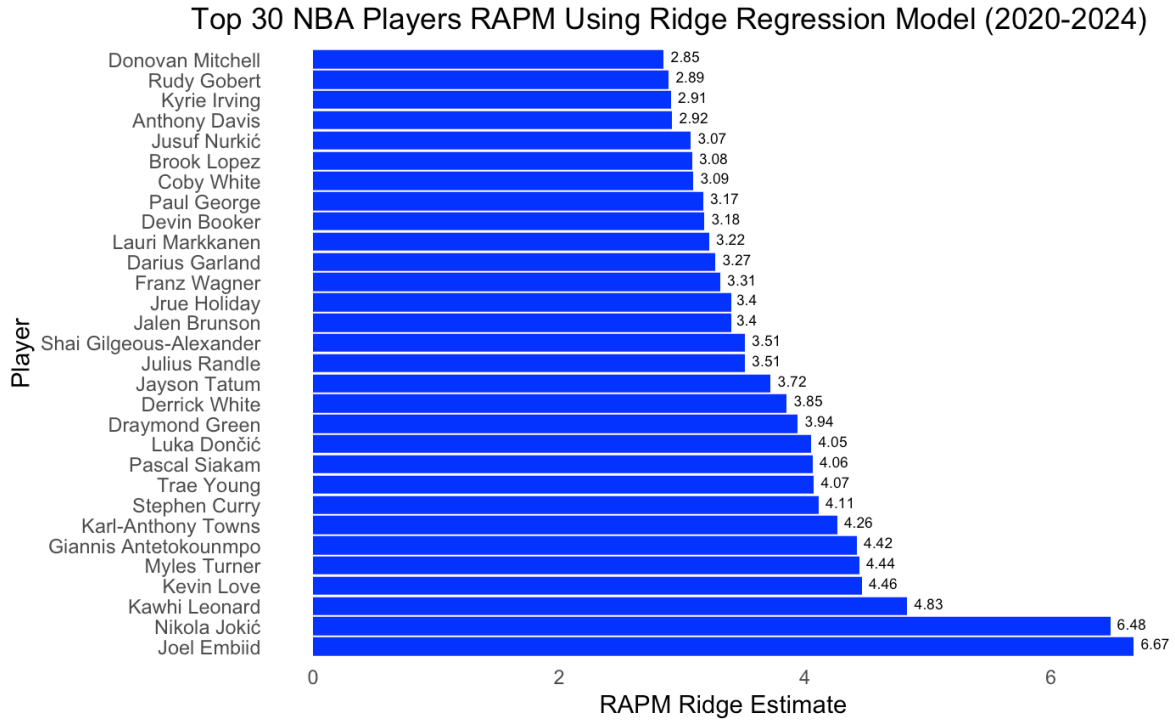


Figure 3: Joel Embiid and Nikola Jokic have the highest RAPM with 6.67 and 6.48, respectively. The 3rd-highest is Kawhi Leonard, with 4.83.

NBA Players Who Played Above Median Minutes

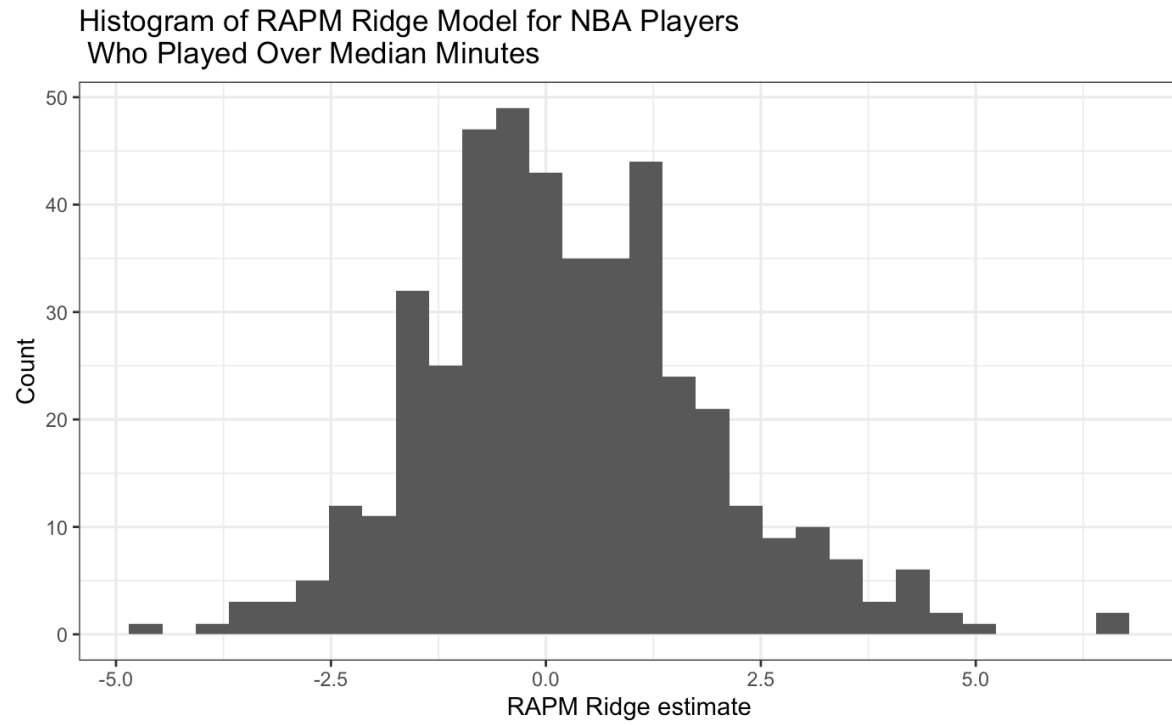


Figure 4: There is wider spread of the coefficients, and less defined bell-shaped distribution.

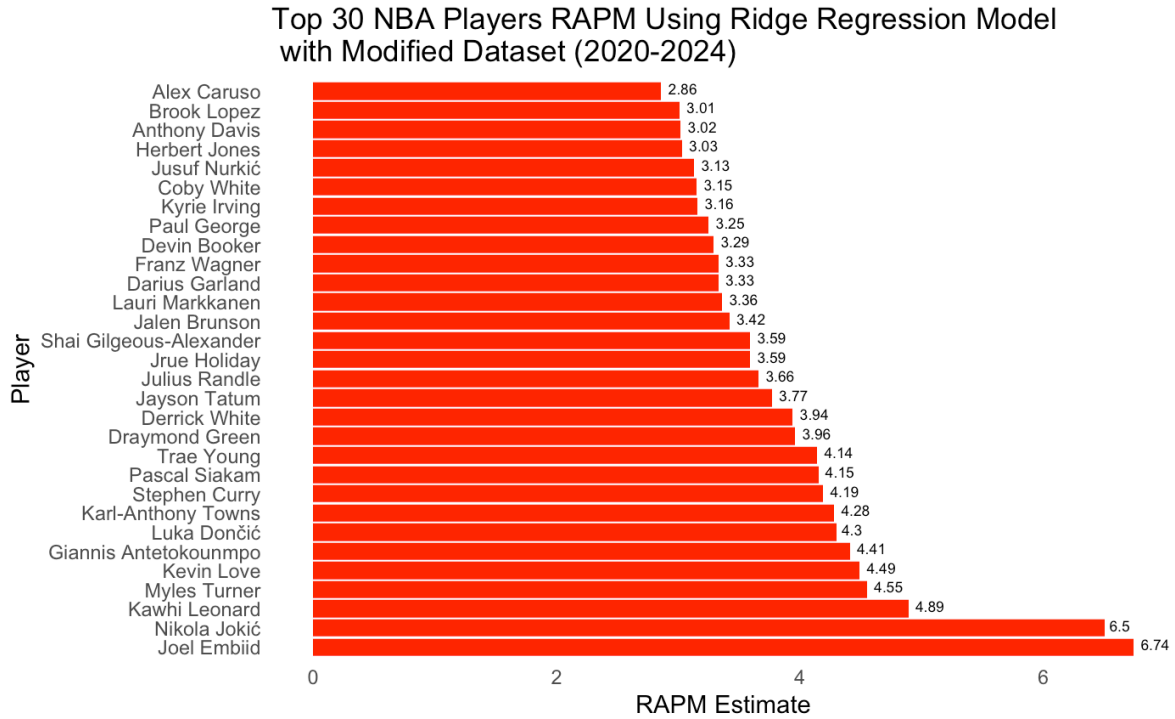


Figure 5: Donovan Mitchell and Rudy Gobert are not in the top 30 with the modified dataset.

Lasso

All NBA Players

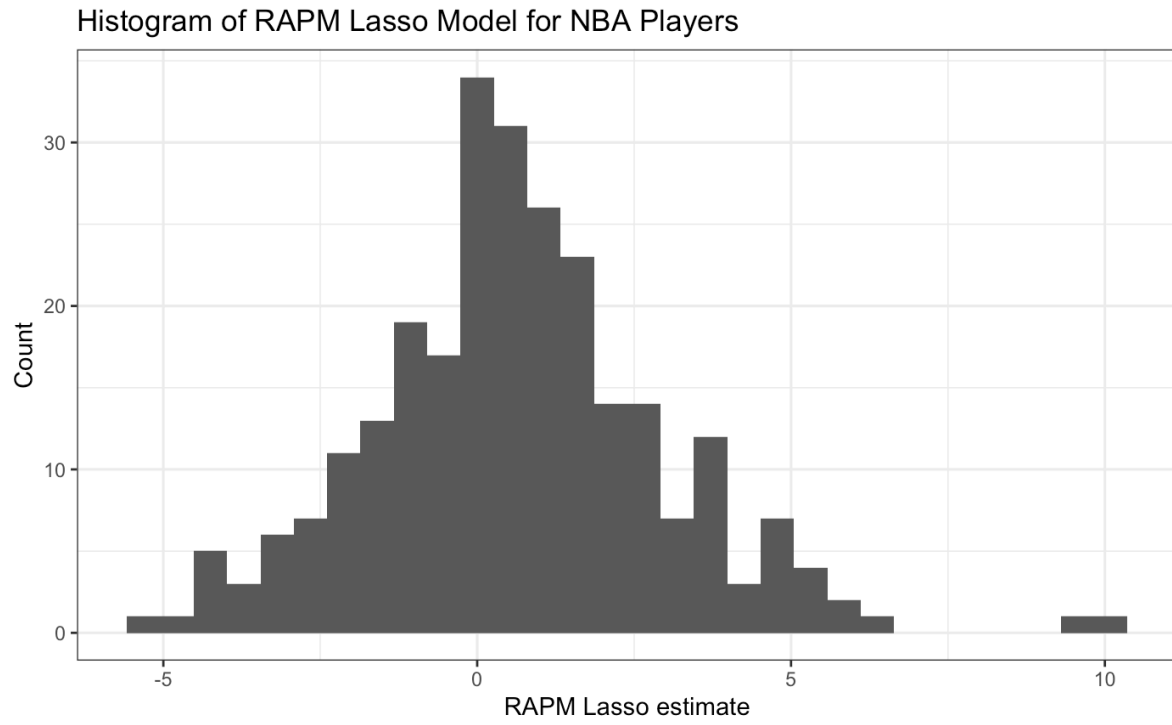


Figure 6: Histogram of Lasso coefficients.

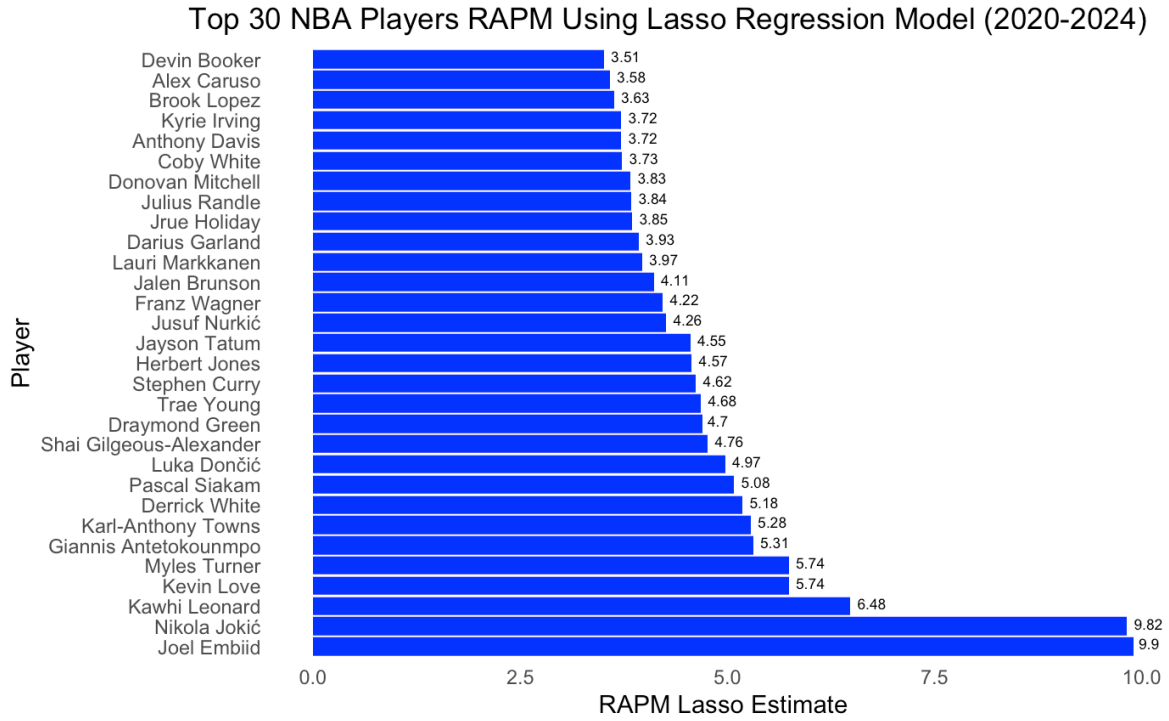


Figure 7: Top 30 players with highest coefficients based on Lasso model.

NBA Players Who Played Above Median Minutes

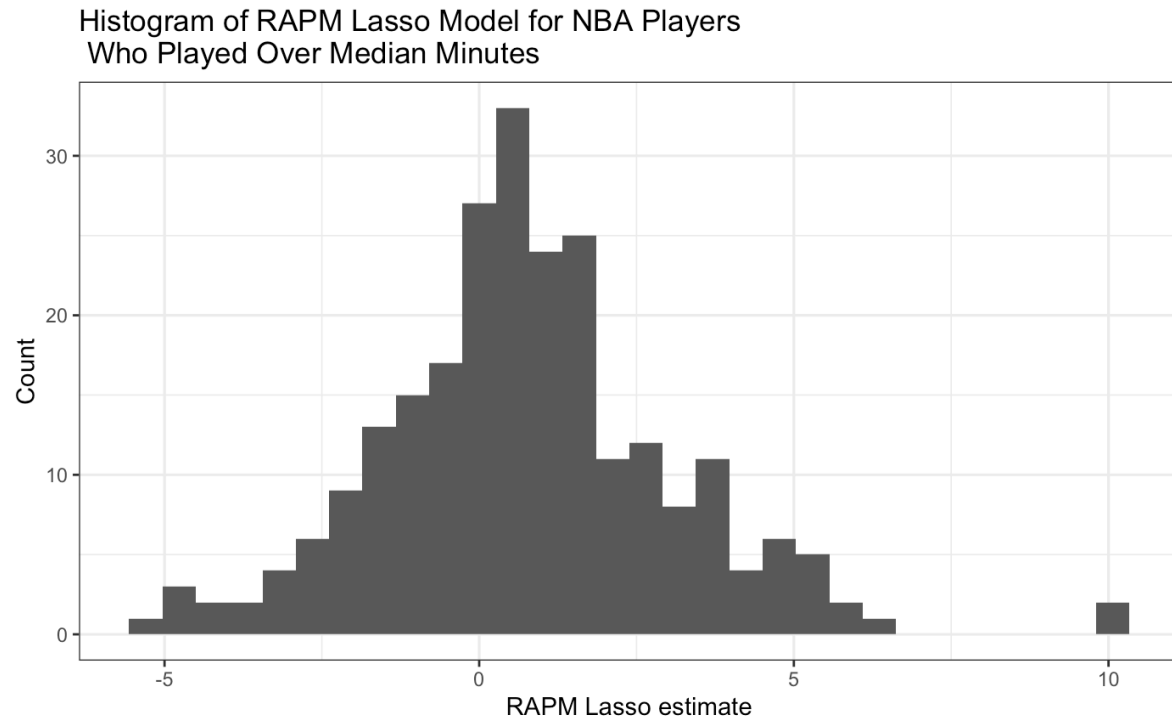


Figure 8: Histogram of coefficients using minutes threshold for Lasso.

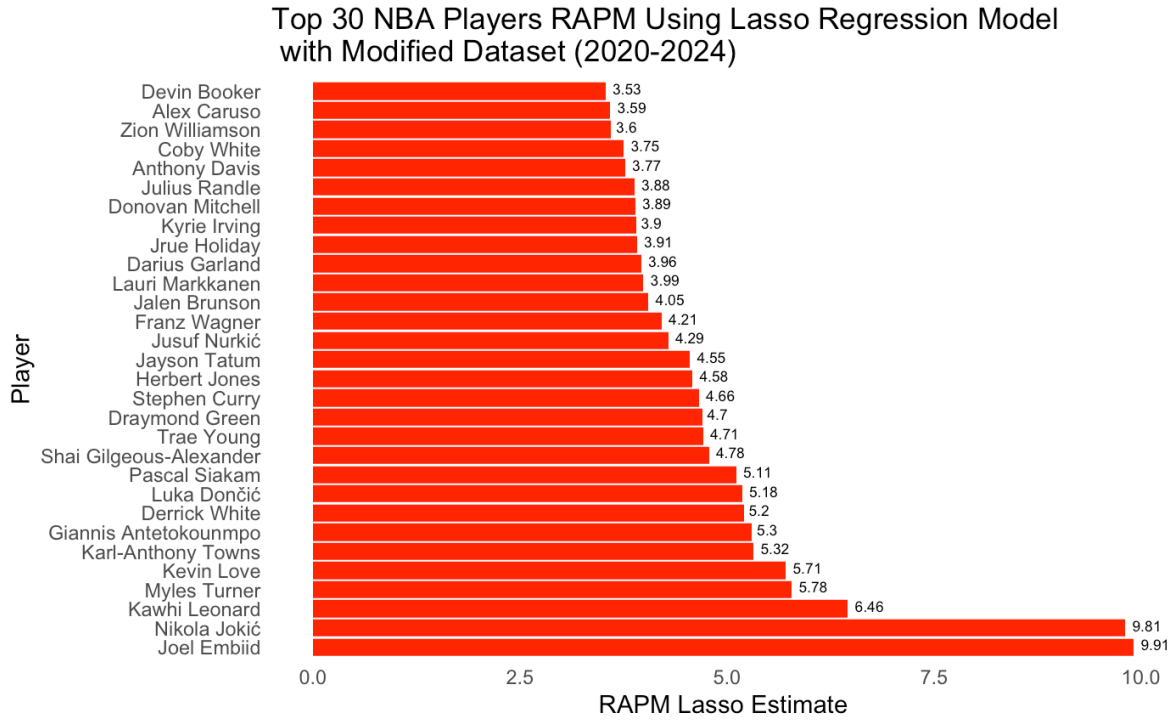


Figure 9: Top 30 players with highest coefficients using Lasso for modified dataset.

Elastic Net

All NBA Players

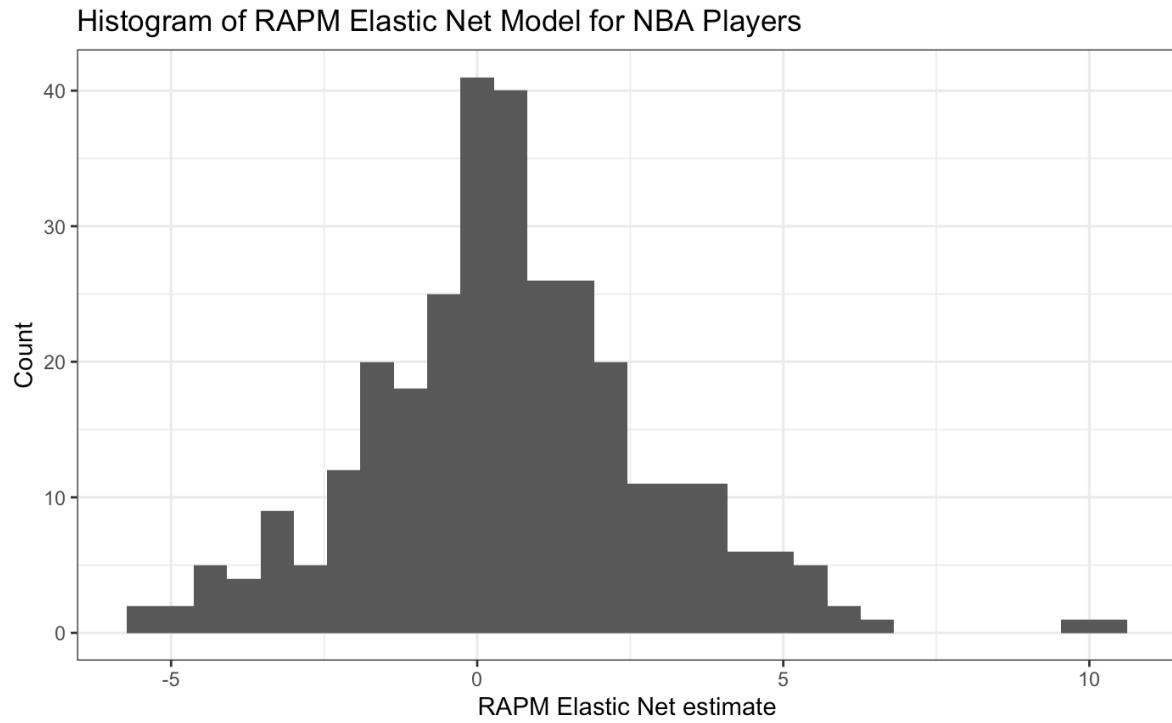


Figure 10: Histogram of Elastic Net coefficients.

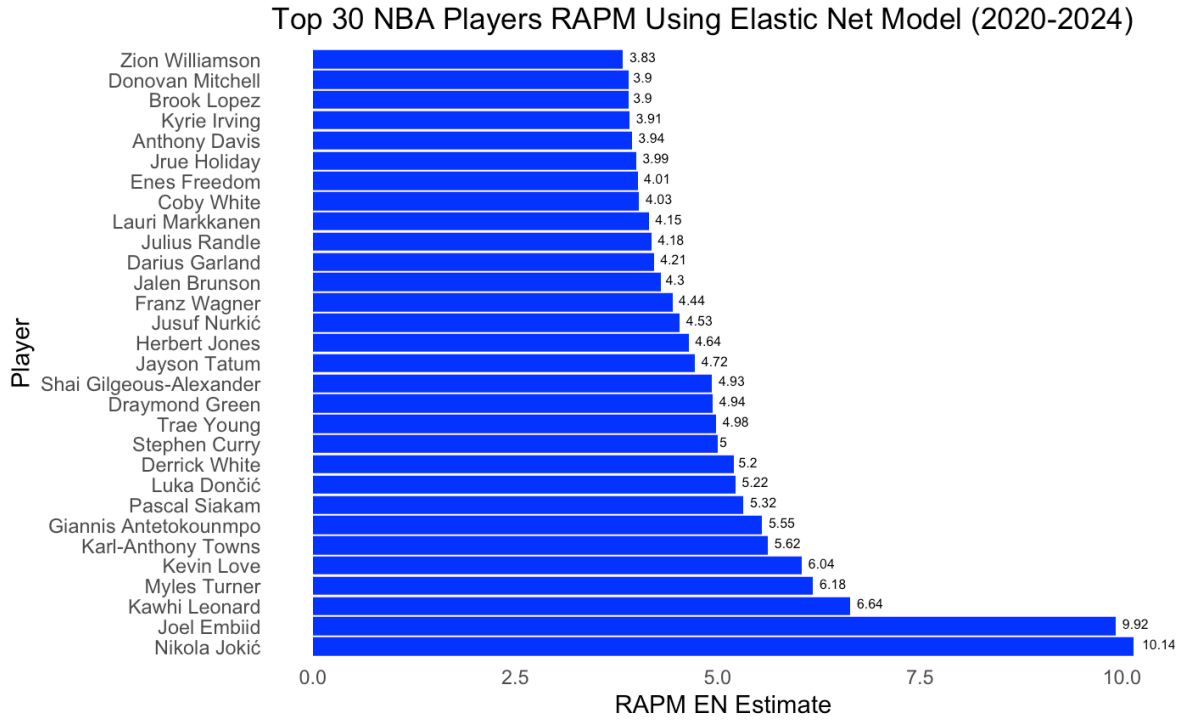


Figure 11: Top 30 players with highest coefficients - Elastic Net model.

NBA Players Who Played Above Median Minutes

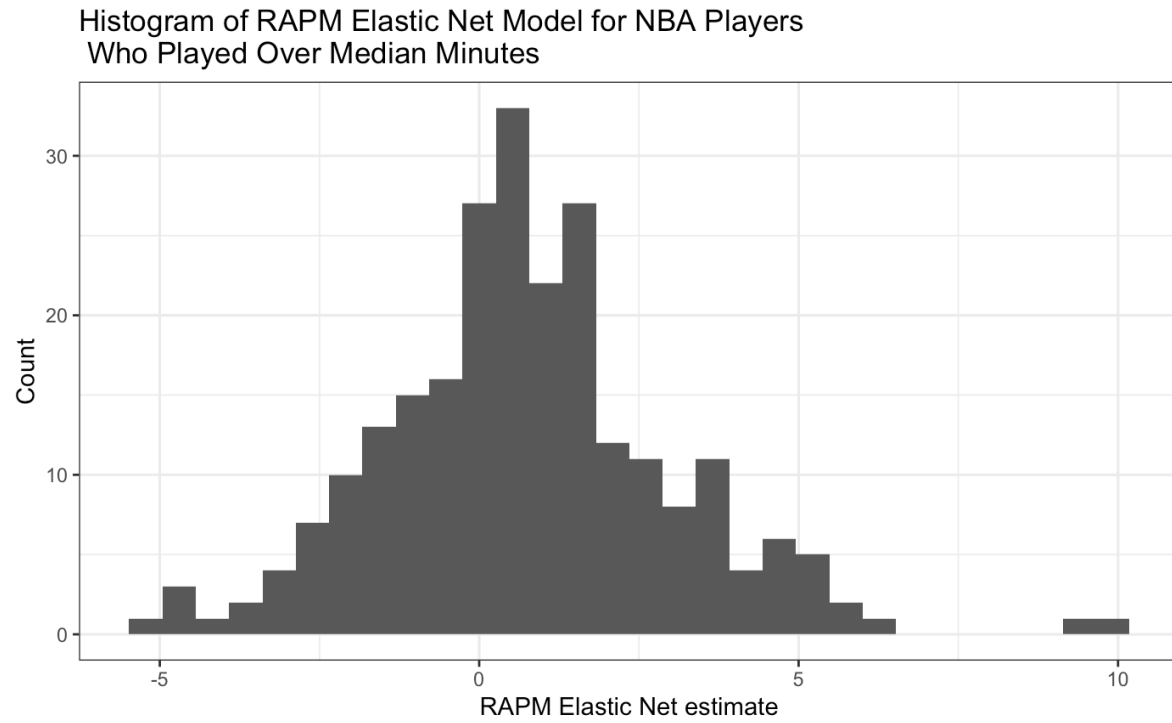


Figure 12: Histogram of Elastic Net coefficients with modified dataset.

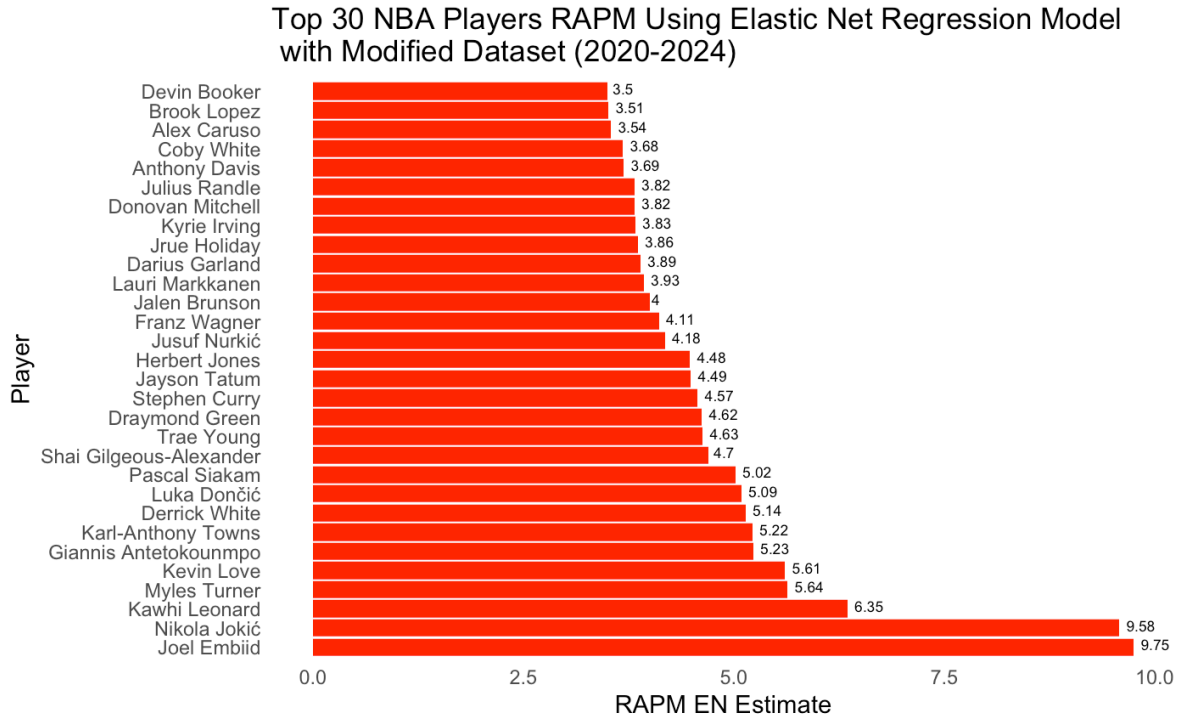


Figure 13: Top 30 players with highest Elastic Net coefficients using modified dataset.

Bayesian Linear Regression

The default chains and iterations used for both Bayesian Linear Regression models were 4 and 2000, respectively.

All NBA Players

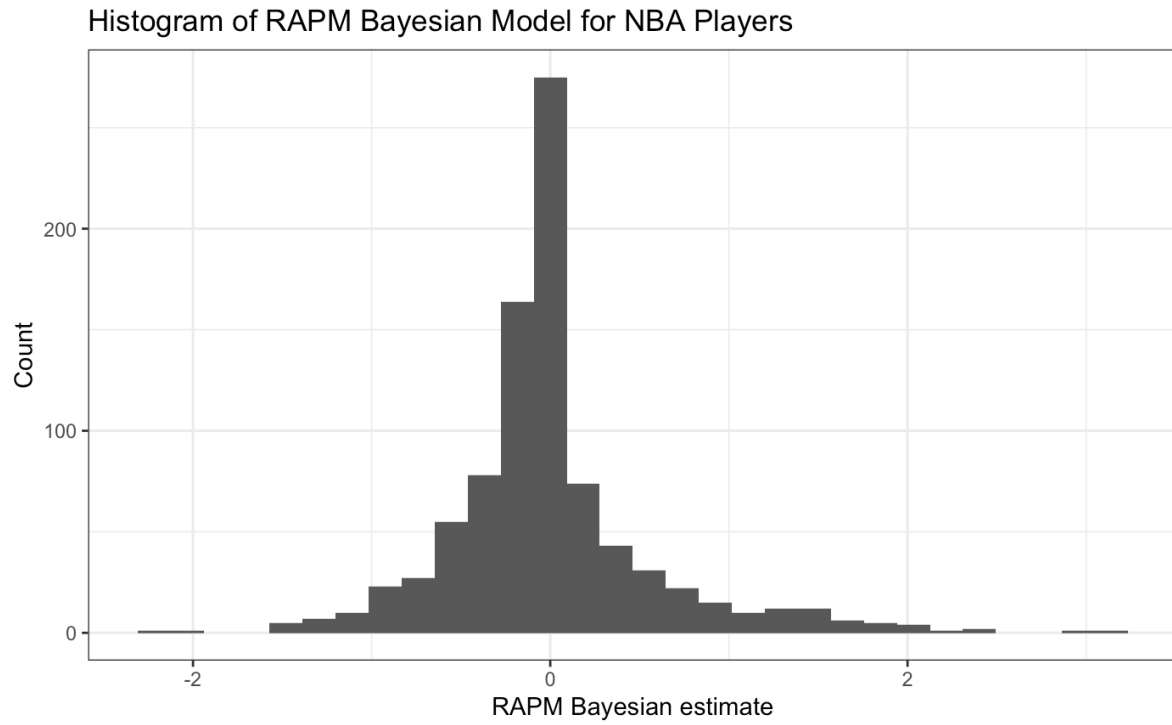


Figure 14: Narrow spread of coefficients suggest low variability.

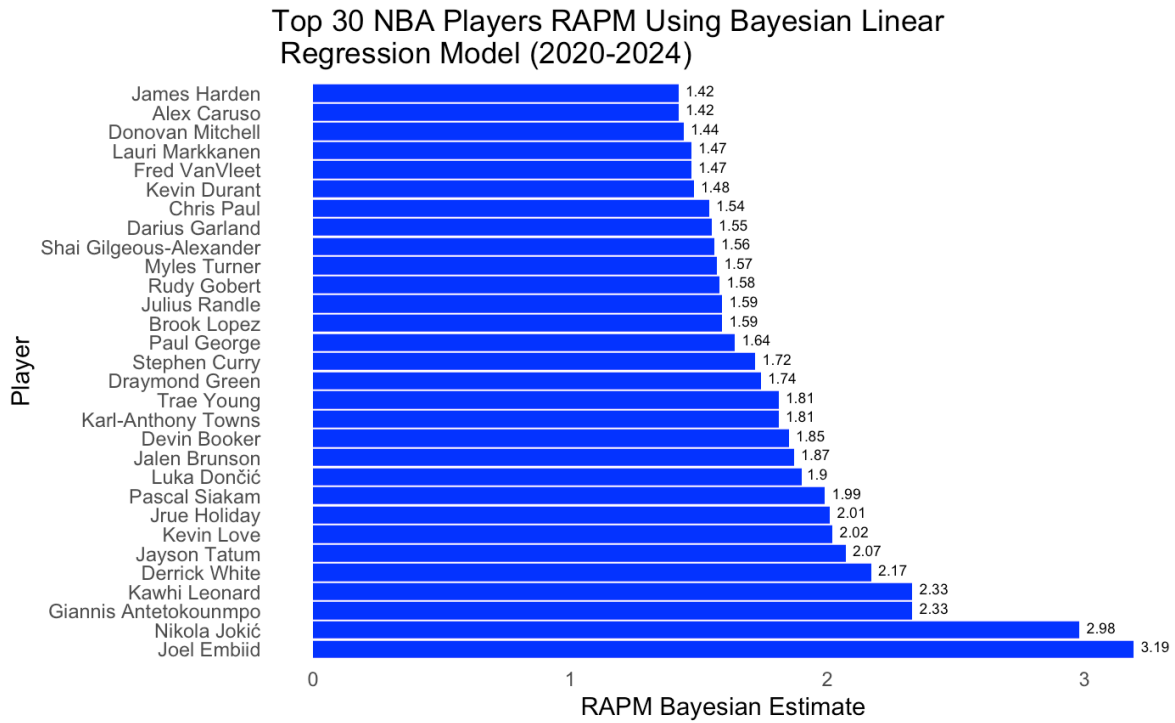


Figure 15: Top 30 players with highest Bayesian coefficients.

NBA Players Who Played Above Median Minutes

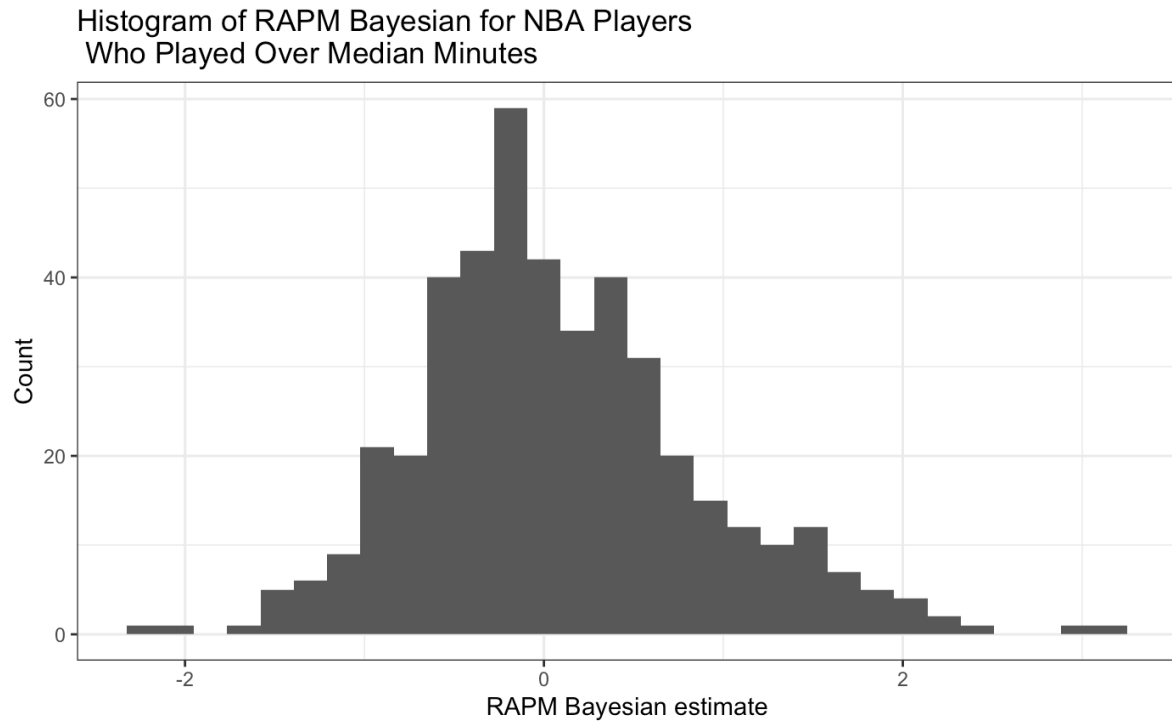


Figure 16: Spread of Bayesian coefficients using modified dataset. Note the right skewness.

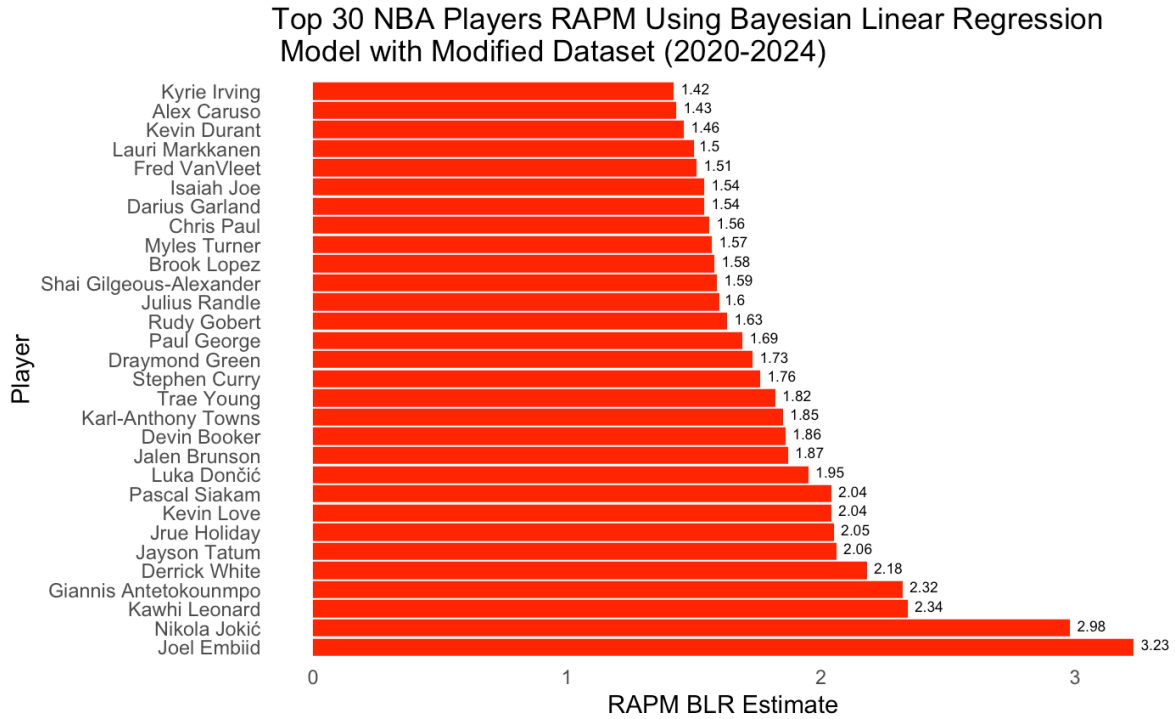


Figure 17: Top 30 player with the highest Bayesian coefficients using modified dataset.

The default chains and iterations used for both Bayesian Linear Regression models were 4 and 2000, respectively.

Metrics for Each Model

	Model	RMSE	R_Squared	Adjusted_R2
1	Ridge Regression	69.868	0.010	0.003
2	Ridge Regression Modified Dataset	69.898	0.010	0.006
3	Lasso Regression	69.899	0.010	0.007
4	Lasso Regression Modified Dataset	69.910	0.009	0.007
5	Elastic Net Regression	69.875	0.010	0.008
6	Elastic Net Regression Modified Dataset	69.912	0.009	0.007
7	Bayesian Linear Regression	70.002	NA	NA
8	Bayesian Linear Regression Modified Dataset	70.013	NA	NA

Figure 18: Ridge Regression produced the lowest RMSE among all models. None of the models produced a high R-squared or Adjusted R-squared.

Due to the computational intensive process for calculating R-squared and Adjusted R-squared, the values for the Bayesian Linear Regression models are not available. However, Bayesian Linear Regression models conducted using 2 as the number of chains with 1000 iterations yielded an R-squared and Adjusted R-squared of 0.003 and -0.004, respectively, for all of the players in the dataset and 0.003 and -0.001 for the subsetting dataset

Discussion

The metrics of the models were not different from one another. The range of Adjusted R-squared values were from 0.003 for the Ridge Regression model including all players from the dataset to 0.008 for the Elastic Net Regression model that included all players. In terms of RMSE, the Ridge Regression model including all of the players had the lowest value with 69.868, while the Bayesian Linear Regression model that included only players who played the median amount of minutes had the highest RMSE with 70.013. Overall each of the models produce a high RMSE and low Adjusted R-squared, which suggests that the models did not perform well in explaining the variability of the response variable (score margin) with the players as the explanatory variables and predicting the actual values. Additionally, the poor performance of the model may indicate that the cross-validation and sampling methods used for the regularization and Bayesian models may not be the best methods for predicting the actual values of the players.

While the metrics of the models indicate low reliability, the models captured well-accomplished players who had the highest coefficients. For instance, the two players with the highest coefficients in each of the models were Joel Embiid and Nikola Jokic, who have both been in the NBA for over 10 years, have won Most Valuable Player awards, and been selected to All-NBA

teams. Jokic has also won a championship. With the exception of the Bayesian Linear Regression model containing all of the players, Kawhi Leonard had the 3rd-highest plus/minus in every other model, which is reasonable considering he has made six All-Star teams, six All-NBA teams, has won 2 championships, and was voted as one of the top 75 NBA players of all-time. Other well-accomplished players such as Stephen Curry, Giannis Antetokounmpo, and Jayson Tatum were also captured in the top 30 for player coefficients for each model.

The player coefficient values varied among the models. Lasso and Elastic Net produced higher coefficient values for the top 30 players. For example, Joel Embiid's coefficient value for Lasso and Elastic Net models that included all players were 9.9 and 9.92, respectively, while his Ridge coefficient value was 6.67 and Bayesian value was 3.19. Because of the different coefficient values produced specifically for the Bayesian models, the ranking of some players fluctuated considerably. In each of the regularization models (Ridge, Lasso, Elastic Net), Myles Turner ranked no lower than 5th among players in the top 30 in coefficient values, higher than Stephen Curry, Jayson Tatum, Shai Gilgeous-Alexander and Luka Doncic. While Turner has been in the league for 10 years and has played mostly as a starter for the Indiana Pacers, he has never averaged more than 18 points per game and was never selected as an All-Star or All-NBA team member like the other guys behind him (Curry, Tatum, Doncic). However, his Bayesian coefficient value rank fell to 20th, with Curry, Tatum, and Doncic ranked ahead of him. This may indicate that the model did a better job of capturing the scoring margin of players when they were on the court than the regularization models.

Another player who was not as well-accomplished as other players in the top 30 but had a high coefficient value in the models was Kevin Love. As with Turner, Love did ranked no lower than 5th in the regularization models. In the Bayesian models, however, his decline in rank was not as precipitous as Turners' was. Love ranked 7th in the Bayesian model including all players, and 8th in the Bayesian model including only players who played the median amount of minutes.

One of the most notable players who did not have the 30 highest coefficients for any of the models was LeBron James. James is considered by many to be one of the top 3 NBA players of all-time, and has played 22 years in the NBA. While he is past his prime, he has still statistically performed very well during the seasons this analysis has used, averaging over 25 points, 7 rebounds and 7 assists with an eFG% of 55 percent. Nevertheless, James' coefficient values may be indicative of the Laker teams he has been on during those four seasons (2020-21 to 2023-24). The win-loss records for those teams were 42-30, 33-49, 43-39, and 47-35, respectively. Therefore, James' coefficient values may be reflective more of the players he played with than his individual abilities to affect the games he played in. Although James' individual statistics indicate he played exceptionally well, his teams did not play well enough to create a scoring margin that may have produced a higher individual coefficient value.

The examples of James, Turner and Love may suggest a couple of inferences about the Plus/Minus stat. First, the teams each player played on may have more of an impact on the player's coefficient values than the player himself. Measuring the impact of a player on team performance is one of the main purposes of the Plus/Minus stat, and it appears that for

these three players the models are applying that concept. A player like James who is among the top players in individual statistical categories may not translate towards a team's overall success. Conversely, if a player isn't among the top players in traditional and advanced stats it doesn't mean he is not contributing to his team's success on the court. Turner and Love played less minutes overall and produced less statistically than James but ranked in the top 10 in the regularization models, indicating their impact on the respective teams whenever they were on the court. Their high coefficient values doesn't mean they are star players on their teams, but they can be identified as solid team contributors that can help a team's success.

Second, while the models did not show a significant drop-off in coefficient values when excluding players who did not play the median amount of minutes, the amount of minutes played could still influence a player's coefficient values. As noted in the previous paragraph, James played more minutes overall (7258.1) than Turner (6041.32) and Love (4037.81). If a player plays a lot of minutes, that means the player plays more possessions and therefore are more opportunities that a team's success or failure can be reflected in the player's plus/minus despite the individual player's statistical achievements. Alternatively, if a player like Turner or Love doesn't play as many minutes as James, that means there are less possessions and opportunities for a team's performance to be reflected in a player's plus/minus valuation. Therefore, there is less data available for those players and may over or underestimate their individual impact on their respective team's performance. It's possible that players like Turner and Love are being used efficiently in the minutes they are playing, resulting in their high coefficient values. It's unclear, however, if their productivity can be representative of a full season of games.

Conclusion

The scope of this analysis compared model performance and player coefficient values produced from Ridge with those produced by Lasso, Elastic Net and Bayesian Linear Regression. Additionally, this paper explores whether utilizing a minutes threshold based on the median amount of minutes played significantly affected player coefficient values. The metrics used to measure the performance of the models (RMSE and Adjusted R-squared) did not provide a statistically significant advantage over Ridge. The minutes limit applied did not significantly alter the coefficient values nor the ranks of the 30 players with the highest coefficient values. However, when comparing player coefficient values between the models, the coefficient values differed. Lasso and Elastic Net produced higher player coefficient values than Ridge, with Elastic Net producing slightly higher coefficients than Lasso. Bayesian produced the lowest player coefficient values among the models, with the highest coefficient player (Embiid) having half the value as the highest coefficient in Ridge models. This may be due to how each of the models handle the amount of features (players) and their corresponding coefficients. More experimentation could be used when applying lambda values for the regularization methods and setting prior parameters for Bayesian Linear Regression, which can help further our understanding for the player coefficient results. Furthermore, applying a train-test split method

instead of cross-validation could improve model performance and provide valuable insights into player coefficient values.

Future extensions presented in this analysis are possible. The models applied can be used to evaluate NBA player production during clutch time, which is defined as a game with a margin of 5 points within the final 5 minutes. An analysis could pinpoint players that perform best during this time period of games and compare a player's regular season and playoff clutch performances to observe if there is an increase or drop-off in productivity. Other types of data that can be used are WNBA regular season as well as Men's and Women's college basketball play-by-play data to evaluate player productivity. Inclusion of other types of player data could enhance our understanding of a player's performance. Dean Oliver's Four Factors, focused on shooting, turnover, offensive rebounding and free throw percentage rates, could be used as input variables to evaluate a player's plus/minus using score margin from play-by-play data. For any development of this analysis, however, data should be aggregated from multiple seasons in order to effectively measure a player's productivity.

References

- "About Box Plus/Minus (BPM)." n.d. *Basketball-Reference.com*. Accessed May 8, 2025. <https://www.basketball-reference.com/about/bpm2.html>.
- "Adjusted Plus-Minus (APM) Explained." 2017. <https://www.nbastuffer.com/analytics101/adjusted-plus-minus/>.
- admin. 2024. "What Is Box Plus/Minus (BPM) in NBA Advanced Stats." *The Midfield*. <https://thefield.com/nba-betting/what-is-box-plus-minus/>.
- Antle, Rob. n.d. "Regularized Adjusted Plus/Minus (RAPM) – Basketball, Stat." Accessed April 4, 2025. <https://basketballstat.home.blog/2019/08/14/regularized-adjusted-plus-minus-rapm/>.
- Bobbitt, Zach. 2021. "When to Use Mean Vs. Median (With Examples)." *Statology*. <https://www.statology.org/when-to-use-mean-vs-median/>.
- "Calculating PER." n.d. *Basketball-Reference.com*. Accessed April 4, 2025. <https://www.basketball-reference.com/about/per.html>.
- "Cross-Validation for Glmnet — Cv.glmnet." n.d. Accessed May 12, 2025. <https://glmnet.stanford.edu/reference/cv.glmnet.html>.
- Damoulaki, Argyro, Ioannis Ntzoufras, and Konstantinos Pelechrinis. 2025. "Lasso Multinomial Performance Indicators for in-Play Basketball Data." *Computational Statistics*, February. <https://doi.org/10.1007/s00180-025-01604-7>.
- Deckard, Dave. 2024. "Demystifying the NBA Plus-Minus Stat." *Blazer's Edge*. <https://www.blazersedge.com/2024/3/31/24117295/nba-plus-minus-portland-trail-blazers-scoot-henderson-record>.
- DeCourcy, Mike. 2024. "The Plus/Minus Basketball Stat: Where It Came from, and Why Experts Vary on How Useful and Informative It Is | Sporting News."

- <https://www.sportingnews.com/us/ncaa-basketball/news/plusminus-basketball-stat-where-it-came-why-experts-vary/fc3e88d011743ba4402ecd1e>.
- “Elastic Net.” n.d. *Corporate Finance Institute*. Accessed April 5, 2025. <https://corporatefinanceinstitute.com/resources/data-science/elastic-net/>.
- Fijal, Damian. n.d. “NBA Net Ratings Throughout the 2023-24 All Star Break. | LinkedIn.” Accessed April 4, 2025. <https://www.linkedin.com/pulse/nba-net-ratings-throughout-2023-24-all-star-break-damian-fijal-xkz9e/>.
- Go, Paolo. 2025. “Box Plus/Minus (BPM) in Basketball: What It Means & How It Works.” *BetUS Sportsbook & Casino*. <https://www.betus.com.pa/nba/news/bpm-basketball-pago/>.
- Jacobs, Justin. 2017. “Relationship Between TS% and eFG%.” *Squared Statistics: Understanding Basketball Analytics*. <https://squared2020.com/2017/10/10/relationship-between-ts-and-efg/>.
- “Lasso Vs Ridge Vs Elastic Net | ML | GeeksforGeeks.” n.d. Accessed May 12, 2025. <https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/>.
- Murel, Jacob, and Eda Kavlakoglu. 2023. “What Is Ridge Regression? | IBM.” <https://www.ibm.com/think/topics/ridge-regression>.
- “NBA Plus-Minus & Impact Metrics in Basketball Explained.” 2017. <https://www.nbastuffer.com/analytics101/nba-plus-minus/>.
- “Ridge Regression Vs Lasso Regression.” n.d. *GeeksforGeeks*. Accessed May 12, 2025. <https://www.geeksforgeeks.org/ridge-regression-vs-lasso-regression/>.
- Sill, Joseph. 2010. “Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing,” March.
- Tim. 2018. “Answer to ”Is Bayesian Ridge Regression Another Name of Bayesian Linear Regression?“” *Cross Validated*. <https://stats.stackexchange.com/a/328345>.
- Van Deventer, Jack. n.d. “The Problems with Plus-Minus and How to Fix Them.” *BasketballScience.Net*. Accessed May 11, 2025. <https://www.basketballscience.net/explanations/the-problems-with-plus-minus-and-how-to-fix-them>.
- “What’s in a Bayesian Model? | R.” n.d. Accessed May 14, 2025. <https://campus.datacamp.com/courses/bayesian-regression-modeling-with-rstanarm/modifying-a-bayesian-model?ex=1>.
- Yurko, Ron, and Quang Nguyen. n.d. “INSTRUCTOR SOLUTIONS: Introduction to Regularized Adjusted Plus-Minus (RAPM) – CMU SCORE Module Preprint Repository.” Accessed April 6, 2025. https://ryurko.github.io/cmu_score_preprints/module_resources/nba-rapm/intro_nba_rapm_SOLUTIONS.html.