Studying Advanced Basketball Metrics with Bayesian Quantile Regression A 3-point Shooting Perspective

Taylor K. Larkin & Denise J. McManus

Information Systems
Statistics and Management Science Department
Culverhouse College of Business
The University of Alabama
USA

Abstract

In light of the recent progression in data collecting methods and the increased emphasis on the 3-point shot in today's NBA, it is desirable to investigate the qualities of good 3-point shooters, especially in regards to more advanced player metrics. Motivated by this opportunity, we implement a regularized Bayesian quantile regression model to identify the most important non-shooting player metrics associated with the best 3-point shooters. The data used for modeling is a combination of SportVU player tracking data and other advanced metrics from the 2015-2016 NBA season. The results are positive and show support that the quantile regression model provides a more comprehensive and accurate assessment of 3-point shooters compared to using Bayesian linear regression. The application of quantile regression models on player tracking data incites opportunities for the development of more advanced analytical models that have the propensity to change the game of basketball.

Keywords: SportVU; regularization; statistical inference; percentile; conditional distribution

1 Introduction

The National Basketball Association (NBA) has been subjected to various trends throughout the years. The current trend in the modern NBA, 3-point shooting, has become an invaluable and dominate tool to win games, significantly changing the game of basketball. In the 1979-1980 season, the season of its inception, NBA teams were shooting less than three 3-point field goals (FGs) per game; however, this number has drastically increased to more than 26 attempts per game in the 2016-2017 season. Additionally, 3-point FG percentage (FG%) has risen from 28% to around 36% in those seasons, respectively (Basketball Reference, 2017). Given the elevated efficiency for these types of shots, increasing the volume of 3-point FGs is advantageous for teams and can make a notable difference in the out- come of a game. Taking into consideration his historic shooting performance in the 2015-2016 NBA regular season, two-time Most Valuable Player Stephen Curry, point guard for the Golden State Warriors, has become the paradigm for demonstrating the advantages of 3-point shooting (Morris, 2015). His record-breaking number of 3-point FGs made (3FGM) played a pivotal role towards his team achieving the record-setting 73 regular season wins.

For this reason, basketball organizations are naturally eager to understand the relevant player metrics corresponding to shooting at long ranges. Traditionally, the most common way to quantify a player's contribution on a team during a game is box score statistics: the number of points scored, number of rebounds, FG%, etc. However, while these are informative at a high-level, they do not capture an entire player's influence towards a team's success or failure. Fortunately, thanks to the SportVU player tracking system (STATS, 2017), more advanced data about a player's spatial characteristics during a game can be captured. This system uses six cameras to track real-time positions and ball activity 25 times per second in each arena, resulting in a wealth of new metrics. Because STATS is an official NBA partner, much of this new information can be found online via the NBA statistics home page (NBA, 2017). The addition of more advanced data collecting entities such as SportVU invites an incredible opportunity for data analysis, moving beyond simply using aggregated box scores.

One of the most popular methods for exploring explanatory relationships in observational data is via linear regression. With this technique, the idea is to find a set of predictor variables that best explain the conditional mean of a response variable. In certain situations, it may be beneficial to explore various parts of the response's distribution, particularly when the extreme values are important, instead of mean behavior alone. In these cases, quantile regression (Koenker and Bassett Jr., 1978) is a useful alternative. The goal of quantile regression is to offer a comprehensive assessment of the effects by estimating at various quantiles (or percentiles) of the response variable. Because different areas of the response can be modeled, it is possible to find predictive relationships at specific quantiles that otherwise would not be useful when modeling the conditional mean (Cade and Noon, 2003).

Quantile regression has been a common approach in several applications related to sports analytics. Hamilton (1997) used censored quantile regression to identify instances of racial discrimination in the NBA during the mid-1990s where traditional regression methods failed. Vincent and Eastman (2009) investigated the effects of various offensive and defensive performance measures of players on their earnings in the National Hockey League, a task deemed too complex for ordinary least squares regression (OLS). Wiseman and Chatterjee (2010) demonstrated that quantile regression outperforms OLS when predicting Major League Baseball player salaries, since the latter is contingent on stricter distributional assumptions. Deutscher, Frick, and Prinz (2013) instituted quantile regression to determine if NBA players are rewarded for good performance in crucial, crunch-time situations, given salary distributions can have amplified amounts of kurtosis.

In each of these studies, quantile regression proved to be a more effective tool than the typical OLS approach for statistical inference. Due to nuances within the sports industry, it is very common for the data to be non-normal and contain outliers, especially when analyzing player salaries. While many of the previous works study impacts on compensation, this method could be used to investigate the effects of advanced player tracking information against specific metrics of interest. Given Curry's monumental impact towards his team via his 3-point shot, one can use quantile regression to identify the most important non-shooting player metrics associated with the *best* 3-point shooters in the NBA, as opposed to trying to find those associated with the *average* 3-point shooter using OLS. Identifying and understanding these characteristics can aid general managers as they try to fill out their rosters. In addition, taking a Bayesian approach in modeling the best 3-point shooters offer a more extensive study of the estimated effects by empirically reflecting their uncertainty. Moreover, introducing regularization into the model presents the opportunity to shrink some of the estimated coefficients towards zero, encouraging sparser and more parsimonious solutions.

The details of this study are presented in the following four sections: Section 2 provides the foundational background about quantile regression and its Bayesian interpretation. Section 3 describes the data and methodology by discussing the basketball metrics (Section 3.1) and the experimental procedures used for creating the model (Section 3.2). Section 4 displays the results of using the proposed approach (Section 4.1) and then shows its advantages over modeling the conditional mean (Section 4.2). Section 5 concludes with a summary and postulates areas for future work.

2 Quantile Regression

In general, estimating the coefficients in quantile regression involves minimizing the sum of asymmetrically weighted absolute residuals (Koenker and Bassett Jr., 1978)

$$\underset{\beta_0,\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i\in\{i:y_i\geq \boldsymbol{x}_i'\boldsymbol{\beta}\}} \tau|y_i-\beta_0-\boldsymbol{x}_i'\boldsymbol{\beta}| + \sum_{i\in\{i:y_i< \boldsymbol{x}_i'\boldsymbol{\beta}\}} (1-\tau)|y_i-\beta_0-\boldsymbol{x}_i'\boldsymbol{\beta}| \qquad (1)$$

for some given quantile level τ where $\boldsymbol{\beta}$ is the $k\times1$ vector of estimated coefficients for τ, β_0 is the intercept, $\boldsymbol{x}_i$ is the predictor matrix of dimension $n\times k$, and y_i is the vector of outcomes of dimension $n\times1$ for n observations and k predictor variables. In this way, different weights are placed on the positive and negative errors of the desired quantile corresponding to under-predicting and over-predicting, respectively. Note that when $\tau = 0.5$ this simply reduces to median regression. Additionally, quantile regression can employ regularization schemes such as lasso, which has been a popular choice due to its sparse nature (Koenker, 2004; Li and Zhu, 2012). More advanced penalties have also been integrated such as the adaptive lasso (Zou, 2006).

Instead of penalizing each coefficient the same as in lasso, the adaptive lasso assigns different weights depending on the coefficient. That is, the objective function becomes

$$\underset{\beta_0,\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i\in\{i:y_i\geq \boldsymbol{x}_i'\boldsymbol{\beta}\}} \tau|y_i-\beta_0-\boldsymbol{x}_i'\boldsymbol{\beta}|+\sum_{i\in\{i:y_i< \boldsymbol{x}_i'\boldsymbol{\beta}\}} (1-\tau)|y_i-\beta_0-\boldsymbol{x}_i'\boldsymbol{\beta}|+\lambda\sum_{j=1}^{k} w_j|\beta_j| \quad (2)$$

where $w_j = 1/|\beta_j|^{\gamma}$, β_j is estimated from an unpenalized quantile regression model, γ is a strictly positive tuning parameter, and $\lambda \geq 0$ controls the amount regularization to introduce. This penalty has been shown to result in better variable selection compared to lasso for penalized quantile regression models (Wu and Liu, 2009).

Quantile regression may also be formulated from a Bayesian perspective. It has been shown that minimizing Equation 1 is equivalent to maximizing the likelihood of a model where the error term follows a skewed Laplace distribution (Yu and Moyeed, 2001; Li et al., 2010; Kozumi and Kobayashi, 2011). This parametrization allows for the quantile regression model to be expressed in the same manner as linear regression, which provides a more efficient way to construct the Gibbs sampler (Alhamzawiet al., 2012). Building on the idea to use a Laplace prior on the coefficients to incorporate the lasso penalty (Li et al., 2010; Alhamzawi, Yu, and Benoit, 2012) extended this specification to formulate the adaptive lasso penalty by placing inverse gamma priors on the respective penalties for each coefficient. Their full Bayesian hierarchical model with the adaptive lasso penalty can be given by the following:

$$\begin{aligned}
&y_i = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + \frac{1-2\tau}{\tau(1-\tau)}z_i + \sqrt{\frac{2}{\tau(1-\tau)}}\xi_i\sqrt{\frac{1}{\sigma}z_i},\\
&p(\beta_0) \propto 1,\\
&p(\xi_i) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\xi_i^2}{2}\right\},\\
&p(z_i|\sigma) = \sigma\exp\{-\sigma z_i\},\\
&p(\beta_j, s_j|\sigma,\lambda_j^2) = \frac{1}{\sqrt{2\pi s_j}}\exp\left\{-\frac{\beta_j^2}{2s_j}\right\}\frac{\sigma}{2\lambda_j^2}\exp\left\{-\frac{\sigma s_j}{2\lambda_j^2}\right\},\\
&p(\lambda_j^2|\delta,t) = \frac{t^{\delta}}{\Gamma(\delta)}(\lambda_j^2)^{-1-\delta}\exp\left\{-\frac{t}{\lambda_j^2}\right\},\\
&p(\sigma) = \sigma^{a-1}\exp\{-b\sigma\},\\
&p(\delta,t) = \frac{1}{t}
\end{aligned} \quad (3)$$

Deriving the posterior distribution yields

$$
\begin{aligned}
&p(\beta_0, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{s}, \sigma, \lambda_1, \ldots, \lambda_k | \boldsymbol{y}, \boldsymbol{X}) \\
&\propto p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{z}, \sigma, \boldsymbol{X}) \prod_{i=1}^{n} p(z_i|\sigma) \times \prod_{j=1}^{k} p(\beta_j, s_j|\sigma, \lambda_j^2) p(\lambda_j^2|\delta, t) p(\sigma) p(\delta, t), \\
&\propto \prod_{i=1}^{n} \frac{\sigma}{\sqrt{\frac{1}{\sigma}\frac{2}{\tau(1-\tau)} z_i}} \exp\left\{ - \frac{\sigma(y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta} - \frac{1-2\tau}{\tau(1-\tau)} z_i)^2}{2\frac{2}{\tau(1-\tau)} z_i} - \sigma z_i \right\} \qquad (4) \\
&\times \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi s_j}} \exp\left\{ - \frac{\beta_j^2}{2s_j} \right\} \frac{\sigma}{2\lambda_j^2} \exp\left\{ - \frac{\sigma s_j}{2\lambda_j^2} \right\} \frac{t^\delta}{\Gamma(\delta)} (\lambda_j^2)^{-1-\delta} \exp\left\{ - \frac{t}{\lambda_j^2} \right\} \\
&\times \sigma^{a-1} \exp\{ - b\sigma \} \frac{1}{t}
\end{aligned}
$$

such that $\boldsymbol{y} = (y_1,...,y_n)$, $\boldsymbol{X} = (x_1,...,x_n)$, $\boldsymbol{\beta} = (\beta_1,...,\beta_k)$, $\boldsymbol{z} = (z_1,...,z_n)$, and $\boldsymbol{s} = (s_1,...,s_k)$. This can be efficiently sampled with the Gibbs sampler described by the authors.

Using this quantile method has several benefits. Unlike the frequentist formulation which typically requires testing on a validation set, the penalty parameters are automatically estimated by the data, since they are treated as unknowns. In addition, instead of having a single point estimate, a Bayesian approach allows for multiple estimates, thereby, delivering a sense of uncertainty about the coefficients. Furthermore, the ability to institute the powerful adaptive lasso penalty within this context can greatly aid in posterior inference at various points in the conditional distribution of the response, especially considering the high dimensionality of player tracking data (Bruce, 2016).

3 Data and Methodology

3.1 Basketball Metrics

Data for this study is based on the 2015-2016 NBA regular season. To prevent biased player estimates, only players who played at least half the season (i.e., at least 41 games) and made at least one 3-point FG are considered. Creating this filter reduced the initial set of 476 players to 296, leaving between nine and 12 players per team. This better signifies those who made an impactful contribution to their respective team throughout the season. All the data for this study can be found via the NBA statistics home page (NBA, 2017).

For the response variable, one could use the number of 3FGM by a player. However, this metric only provides a summed total and does not incorporate the difficulty of the 3FGM. Naturally, better 3-point shooters will make more challenging 3-point FGs. Since the goal is to find the qualities of the best 3-point shooters, this work derives a new metric, 3-point rating (3PR), that weights more laborious 3FGM (i.e., when a defender is close to the shooter) more and easier ones less (i.e., when the shooter is wide open). Breaking down 3FGM at varying defender distances, 3PR can be created by

$$3PR = 0.4*3FGM_{0-2} + 0.3*3FGM_{2-4} + 0.2*3FGM_{4-6} + 0.1*3FGM_{6+} \qquad (5)$$

such that the respective subscript represents the closest defender distance range in feet (e.g., $3FGM_{0-2}$ = number of 3FGM when closest defender is 0-2 feet away). Note that the weights are a convex combination to reflect more emphasis on 3FGM when the nearest defender is very close. Modeling this instead of just 3FGM will better emphasize the qualities of good 3-point shooters.

To analyze the effects of non-shooting player tracking metrics on 3PR, only certain metrics are extracted from the player tracking page. These are grouped into four different categories specific to this study: defense, possession, passing, and agility. In addition to these metrics of interest, several others are included to serve as controls (e.g., what team a player is on, what positions he plays, etc.). It is necessary to estimate these in the model to help control for as much variability in 3-point shooting as possible (i.e., did Stephen Curry make a large number of 3-point shots because he played the guard position on a good team?).

The complete list of 54 metrics used in this study is given in Table 1. These are all per game statistics where applicable or unless otherwise noted.

Table 1: List of the 54 metrics used in the modeling. For space purposes, bold numbers indicate how many metrics are contained in each definition.

Metric	Description	Role
3PR	Rating to account for the difficulty of 3FGM	Response
Steals/Blocks/Defensive Rebounds	Number of steals/blocks/defensive rebounds (**3**)	Defense
Defensive Field Goal %	FG% of the opponent while the player was defending the rim	Defense
Post/Elbow Touches	Number of touches made by a player near the post/elbow (**2**)	Possession
Points/Outside Touch	Number of points scored per touch not at the paint, post, or elbow	Possession
Average Dribble/Touch	Average number of dribbles per touch	Possession
Assist Points/Assist	Number of assist points created per assist	Passing
Secondary Assist	Number of passes made by a player who earned an assist on a made shot	Passing
Assist Efficiency	Percentage of potential assists that became assists	Passing
Drives	Number of times a player starts at least 20 feet of the hoop and dribbles within 10 feet of the hoop, excluding fast breaks	Passing
Distance Traveled	Distance run by player measured in miles	Agility
Average Rebound Distance	Average distance of a rebound from the goal	Agility
Average Offensive/Defensive Speed	Average speed in miles per hour of all movement (sprinting, jogging, standing, walking) by a player while on the court on offense/defense (**2**)	Agility
Games Played	Number of games played	Control
Position	Binary denoting player position (guard, forward, and/or center) (**3**)	Control
Wins	Number of wins accrued by team	Control
Usage %	Percentage of team plays used by a player when he is on the floor	Control
PIE Difference	Difference between a player's player impact estimate (PIE) and the team's average PIE—PIE measures the overall statistical contribution against the total statistics in a game	Control
Pace	Represents the number of possessions per 48 minutes for a player	Control
Team	Team binary (**29** where all zeros represents the 30th team)	Control

The majority of metrics included are directly from the NBA statistics website. However, in order to achieve diversity in the player tracking metrics and decrease high pairwise correlations, some new metrics are created. For instance, *Assist Efficiency* is calculated by dividing the number of assists by potential assists. Potential assists are defined as passes made that would have led to an assist, despite whether the teammate made the FG. *Assist Efficiency* can be considered a counterpart to FG%. In this way, a player's ability to set their teammates up for easier FGs can be considered. Moreover, *Assist Point/Assist* is simply the number of assist points created divided by the number of assists. This allows a comparison amongst players whose assists primarily originate from behind the 3-point arc or closer to the goal. Finally, *Point/Outside Touch* is tabulated via the following formula:

$$\text{Points/Outside Touch} = \frac{T * P_T - (A * P_A + O * P_O + E * P_E)}{T - (A + O + E)} \qquad (6)$$

such that *T*, *A*, *O*, and *E* represent the number of per game touches, paint touches, post touches, and elbow touches, respectively, and *P* symbolizes the number of points scored at the respective subscript (e.g., P_A = points scored in the paint). This calculation gives an approximated estimate for how many points are scored further away from the goal per touch, which is expected to be positively related to 3PR. This is evident by the fact that this metric has the highest degree of linear association with 3PR among all the metrics with a correlation coefficient of $r = 0.60$.

Along with the number of games played, position, and team, other metrics are included as controls. *Wins* account for the success of a player's team. *Usage %* and *PACE* account for how much a player uses the possessions he is given. *PIE Difference* attempts to account for teammate interactions by measuring relative performance (i.e., does a player's performance far exceed that of his teammates?). Adding these types of control variables will help deliver a more accurate estimation of the effects of interest.

3.2 Experimental procedure

To implement the Bayesian quantile regression model with the adaptive lasso penalty (BALQR), the **bayesQR** package (Benoit and Van den Poel, 2017) within the R environment for statistical computing version 3.3.3 (R Core Team, 2017) is used for analysis. The number of Markov Chain Monte Carlo (MCMC) draws is set to 500,000 with the first 20% discarded from posterior inference as burn-in rounds. A large number of draws is necessary to achieve fairly stable estimates given the data are moderately correlated. Visual inspection of the MCMC chains is used to ensure adequate convergence. To analyze the behavior of each metric of interest, we implement quantile process regression. That is, separate quantile regression models are executed at various points in the conditional distribution of the response. For this study, a model is constructed for $\tau = \{0.05, 0.25, 0.50, 0.75, 0.95\}$ on 3PR. In addition, prior to the modeling, the non-binary metrics are standardized to have a mean of zero and standard deviation of one to more fairly compare the effects.

As noted by Deshpande and Jensen (2016), who use Bayesian linear regression with a lasso penalty to estimate NBA player impact on win probability, some strong assumptions exist when using linear regression. In their work, they specifically mention the independence of observations and that the errors should have constant variance. Here, the former assumption is reasonably relaxed since only one season of data is considered (i.e., each observation represents a different player). For the latter, since quantile regression does not assume the errors exhibit constant variance, this assumption may also be circumvented. The reduction in assumptions further adds to the benefits of making inference using a more robust method like quantile regression (Wiseman and Chatterjee, 2010). To benchmark the performance of using quantile regression compared to estimating the conditional mean, Bayesian linear regression with a lasso penalty (Blasso) is also implemented within R using the **monomvn** package (Gramacy, 2017). For a fair comparison, the same number of draws and burn-ins are executed. Note that no normalization is done via the *blasso* function since the data is already standardized. All other arguments are left at the function's default settings.

4 Results and discussion

4.1 Quantile process regression

Figure 1 displays the quantile process regression plots. By analyzing these plots, a more extensive picture of the changing coefficient estimates can be made, resulting in some natural and interesting inferences. Note that the closer the coefficient estimate is to zero, the less impact it has, according to the model.

Points/Outside Touch, *Distance Traveled*, *Average Rebound Distance*, and *Average Offensive Speed* all have mainly positive coefficients while *Blocks*, *Post Touches*, *Elbow Touches*, and *Drives* are all mainly negative. The sign of these effects are undoubtedly a product of their role on the floor. Top-tier shooters like Curry will drive the ball less and typically possess the ball more away from the goal as they facilitate the offense, leading to increased speed and distance traveled on the outer edges of the court as opposed to at the elbow or the post. The positive effects of *Average Rebound Distance* could also explain the equally negative effects for *Blocks*, given the increased distance from the goal. As expected, the effects for *Points/Outside Touch* are positive since it is directly tied with shooting at longer ranges.

While some metrics show primarily positive or negative values, others show different effects depending on the quantile. *Steals* does not appear to be very useful in the lower quantiles, but shows a positive impact in the upper ones. This is consistent with the advantage that quantile regression can uncover different relationships depending on the quantile. *Defensive Field Goal %*, *Assist Efficiency*, *Assist Points/Assist*, and *Defensive Rebounds* fluctuate around and close to zero above and below depending on the quantile, denoting that these are probably not a large factor in determining 3PR. The higher values of *Steals* for good 3-point shooters is reasonable since they are also commonly a primary ball-handler and, thereby, guarding other ball-handlers. Although the overall impact of *Defensive Field Goal %* is marginal, the negative effect in the lower percentiles also is logical since this denotes how well a player defends at the rim. These types of players typically reside at the center and forward positions and may not shoot many 3-point FGs.

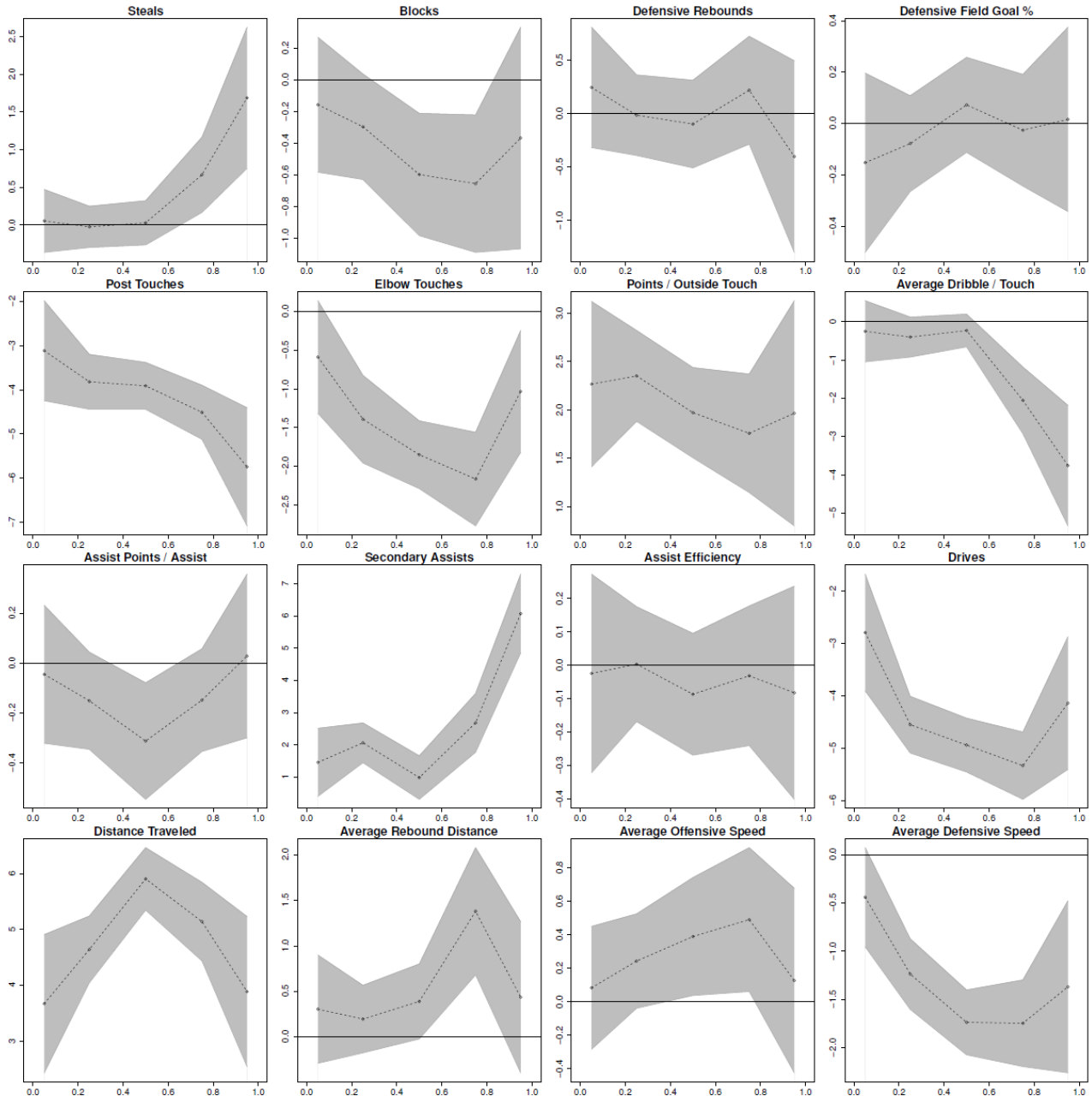

Figure 1: Quantile process regression plots with +/- 1 standard deviation around the posterior means (shaded region). The black line denotes the zero axis to display the shrinkage towards zero from the adaptive lasso penalty.

Other metrics present intriguing patterns. *Secondary Assists* presents a strong, positive effect for the upper quantiles. This can again be attributed to good 3-point shooters also being good facilitators of the offense, involving their teammates by passing the ball around frequently. *Average Defensive Speed* generally trends downward, meaning that increases in speed while on defense can negatively impact 3-point shooting. This could be indicative of those shooters who are known for being pure offensive players like James Harden, who achieved the second largest 3PR, as opposed to making the most effort on defense.

Perhaps the most interesting finding from Figure 1 is the behavior of *Average Dribble/Touch*. This metric has a larger, negative relationship and decreases as the quantiles increase closer to one. This indicates that great 3-point shooters do not hold the ball for relatively long periods of time when controlling for the other metrics.

More than likely, this can be explained by the concept of catch-and-shoot 3-point FGs (i.e., 3-point FGs immediately following a reception of a pass). Therefore, defenses could focus on forcing Curry and other great shooters to take isolated 3-point shots to try to reduce their success, which was demonstrated in the 2015-2016 NBA Finals (Ruiz, 2016). These types of conclusions are possible with the help of quantile process regression. Because this approach provides a more complete picture of how these metrics interact with one another, general managers can profile role players, or even college prospects, based on their advanced statistics using these insights to indicate who can be successful 3-point shooters.

4.2 Comparison to Bayesian linear regression

As done in previous studies, it is noteworthy to benchmark the performance of quantile regression compared to simply modeling the conditional mean. Figure 2 displays the posterior means and standard deviations for each player tracking metric of interest for BALQR at the 95^{th} percentile and Blasso.

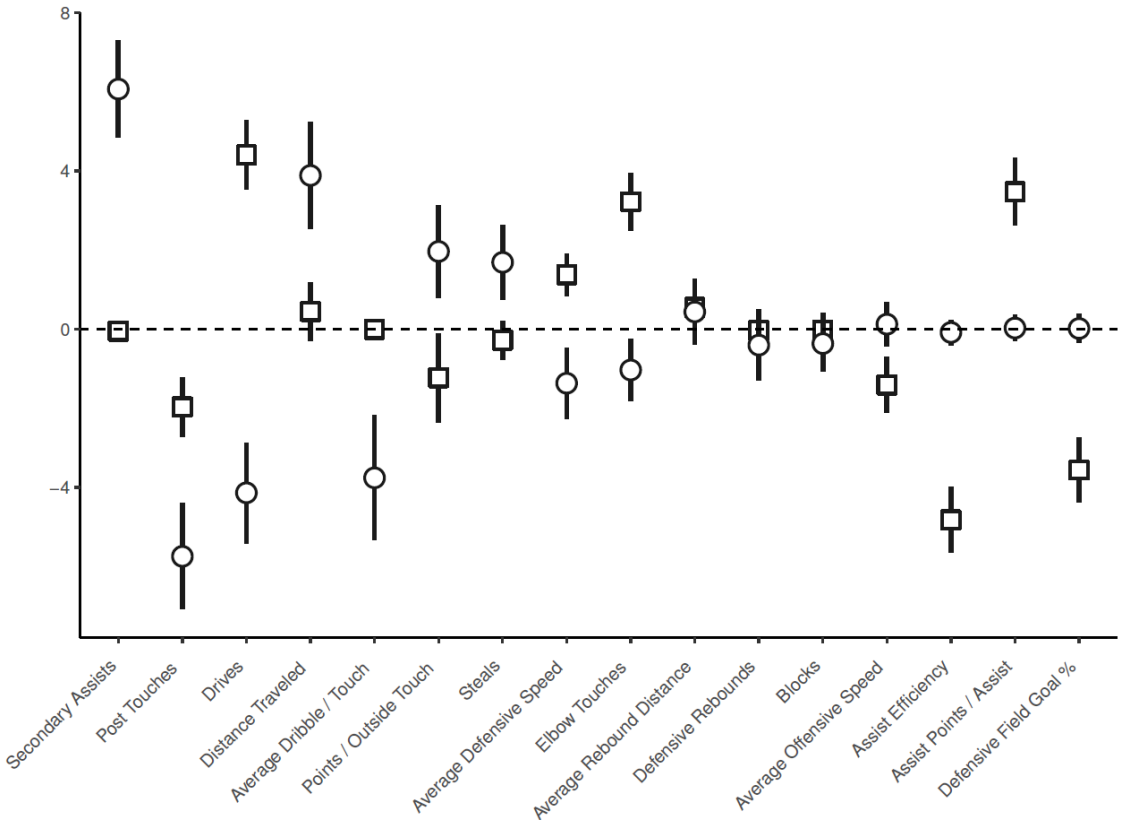

Figure 2: Posterior means with +/- 1 standard deviation for BALQR at the 95^{th} percentile (circles) and Blasso (squares). These are ranked from those with the highest coefficient in magnitude (left) to the lowest (right) according the BALQR. The zero axis is again denoted.

Based on the magnitude of the posterior means, the most important metrics to consider for the best 3-point shooters are *Secondary Assists*, *Post Touches*, *Drives*, *Distance Traveled*, and *Average Dribble/Touch*. Of these top metrics, Blasso presents the opposite effect for *Drives* and deems *Secondary Assists* and *Average Dribbles/Touch* as not very useful. Moreover, Blasso assesses a negative effect for *Points/Outside Touch*, which is unlikely given that this particular metric has the most positive linear association with 3PR of any metric, including the controls. It is these types of contradictory conclusions that make the case for using quantile regression when analyzing the most impactful player tracking metrics for the *best* 3-point shooters, rather than subjecting this inference to strictly mean-centered models. The advantages can be further demonstrated by evaluating the fitted values for the best shooters with each method. Table 2 lists such values for BALQR and Blasso calculated from their posterior means.

Table 2: Fitted values for players in the top 5% of 3PR. Bold and italics indicate the method closest to the actual value.

Player	3PR	BALQR	Blasso
Stephen Curry	81.60	***81.89***	46.60
James Harden	55.00	***55.51***	34.70
Klay Thompson	53.50	***54.67***	33.56
Damian Lillard	47.40	***45.61***	26.80
Kyle Lowry	43.20	***43.16***	22.55
Kevin Durant	41.80	***44.32***	30.58
Paul George	40.50	***43.56***	29.21
J.R. Smith	38.00	***38.33***	25.41
J.J. Redick	37.90	***42.22***	33.56
C.J. McCollum	36.40	48.88	***29.43***
Kemba Walker	35.90	48.44	***30.46***
Mirza Teletovic	35.40	***33.20***	18.98
Isaiah Thomas	35.40	***39.41***	21.89
Robert Covington	35.30	***35.93***	20.04
Isaiah Canaan	34.50	***35.96***	22.16

For all but two players, BALQR better estimates 3PR. Note, however, that the in-sample predictions from BALQR will suffer for the middle and lower percentile of shooters, since these coefficients are based on the 95^{th} percentile. If one was interested in studying other percentiles, then he or she should calculate the coefficient estimates at the desired quantile level. In summary, for the best 3-point shooters, quantile regression promotes a more accurate assessment of those relevant, non-shooting player tracking metrics compared to using its linear counterpart.

5 Conclusions and future work

In this study, we use a regularized Bayesian quantile regression model to investigate the most important non-shooting player metrics associated with the best 3-point shooters in the NBA. Using quantile process regression, we found that the effects differed depending upon the quantile level, leading to more comprehensive conclusions. The most important metrics for the best 3-point shooters consisted of *Secondary Assists*, *Post Touches*, *Drives*, *Distance Traveled*, and *Average Dribble/Touch*. These metrics, in particular *Average Dribble/Touch*, present an interesting combination of advanced statistics for general management to understand about the characteristics of good 3-point shooters, which can be used to evaluate potential players. Using this quantile approach is shown to lead to more sound inference in the presence of the control variables compared to modeling the conditional mean.

While this study presents noteworthy and practical results, it is, however, based upon one season of data. Including more seasons may yield some differences, though the independence assumption would be violated. In addition, inspection of the trace plots from BALQR indicates that a few of the metrics may have had some difficulty converging at some of the quantiles. This is likely due to the correlated nature of the data, making the mixing of the chains sometimes challenging. Future work could be to improve the convergence via transformations or using different priors. Regardless, the estimates appear to have reasonable convergence for this study. In addition, some of the effects may be inflated by role players playing during periods of a game where the outcome has virtually been decided (i.e., garbage time). Perhaps analyzing these metrics as a function of a player's impact, similar to the work of Deshpande and Jensen (2016), would be an intriguing endeavor.

Furthermore, the findings of this research can be the groundwork for developing additional statistical models based upon modeling other metrics such as PIE or a more complex weighting of 3FGM. The continued advancement of systems like STATS SportVU will permit researchers to collect and analyze additional player metrics on the court. Moreover, this analysis can be used to investigate other kinds of player information from avenues such as wearable technologies or even personality tests. Many corporations use personality tests to find the best fit for the organization and corporate team; hence, these types of tests could add another dimension to understanding sports players and team dynamics. Modeling the more intangible aspects in this way can unlock truly revolutionary discoveries that will continue to change the game.

Our goal is that the approach described in this work can aid in understanding the relationships amongst the plethora of advanced metrics and can ultimately lead to improving overall team performance.

References

Alhamzawi, R., Yu, K., and Benoit, D. F. (2012). Bayesian adaptive lasso quantile regression. *Statistical Modelling*, 12(3):279–297.

Basketball Reference (2017). NBA league averages. Available at http://www.basketball-reference.com/leagues/NBAstats.html. Accessed 2017-03-23.

Benoit, D. F. and Van den Poel, D. (2017). bayesQR: A Bayesian approach to quantile regression. *Journal of Statistical Software*, 76(1):1–32.

Bruce, S. (2016). A scalable framework for NBA player and team comparisons using player tracking data. *Journal of Sports Analytics*, 2(2):107–119.

Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420.

Deshpande, S. K. and Jensen, S. T. (2016). Estimating an NBA player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, 12(2):51–72.

Deutscher, C., Frick, B., and Prinz, J. (2013). Performance under pressure: Estimating the returns to mental strength in professional basketball. *European Sport Management Quarterly*, 13(2):216–231.

Gramacy, R. B. (2017). *monomvn: Estimation for multivariate normal and Student-t data with monotone missingness*. R package version 1.9-7.

Hamilton, B. H. (1997). Racial discrimination and professional basketball salaries in the 1990s. *Applied Economics*, 29(3):287–296.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89.

Koenker, R. and Bassett Jr., G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81(11):1565–1578.

Li, Q., Xi, R., and Lin, N. (2010). Bayesian regularized quantile regression. *Bayesian Analysis*, 5(3):533–556.

Li, Y. and Zhu, J. (2012). L1-norm quantile regression. *Journal of Computational and Graphical Statistics*.

Morris, B. (2015). Stephen curry is the revolution. Available at http://fivethirtyeight.com/features /stephen-curry-is-the-revolution/. Accessed 2017-03-23.

NBA (2017). Stats home. Available at http://stats.nba.com/. Accessed 2017- 12-14.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruiz, S. (2016). What went wrong for Steph Curry in the 2016 NBA finals? Available at http://ftw.usatoday.com/2016/06/stephen-curry-2016-nba-finals-stats-golden-state-warriors. Accessed 2017-03-23.

STATS (2017). STATS SportVU. Available at http://www.stats.com/sportvu-basketball-media/. Accessed 2017-03-23.

Vincent, C. and Eastman, B. (2009). Determinants of pay in the NHL: a quantile regression approach. *Journal of Sports Economics*.

Wiseman, F. and Chatterjee, S. (2010). Negotiating salaries through quantile regression. *Journal of Quantitative Analysis in Sports*, 6(1).

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, pages 801–817.

Yu, K. and Moyeed, R.A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*,54(4):437–447.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.