MARCH MADNESS PREDICTION USING QUANTILE REGRESSION

by

Kimberly Mays

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Mathematics

Charlotte

2022

Approved by:

_____

Dr. Eliana Christou

_____

Dr. Michael Grabchak

_____

Dr. Shayou Li

ABSTRACT

KIMBERLY MAYS. March madness prediction using quantile regression. (Under the direction of DR. ELIANA CHRISTOU)

The annual NCAA Division I men's basketball tournament is one of the most prominent sporting events in the United States and the subject of much interest among fans and statisticians alike. In this work, we propose a new method for predicting the results of the tournament based on a semiparametric quantile regression model. The idea is to estimate a win probability by averaging across multiple conditional quantiles. To demonstrate the finite sample performance of the proposed methodology, we predict the tournament results for the years 2016 - 2019 and 2022. The results are then compared with other commonly used methods and rankings. Our method is competitive and offers a novel approach for use in bracket prediction.

ACKNOWLEDGEMENTS

I have been helped by a number of people in the process of completing this project. Thank you! Special thanks are due to Dr. Eliana Christou for her untiring advice and encouragement. She was willing to take on both a new student and a completely new area of research, for which I am very grateful. Even more, I left our meetings more confident and focused, which is a mark of a great advisor and mentor.

Dr. Shayou Li and Dr. Michael Grabchak also warrant special thanks for their willingness to serve on my committee and for the impact they each had on my progress in the classroom. Dr. Li's experimental design course and Dr. Grabchak's applied probability course were two of my favorite courses. They were also kind and helpful in their encouragement and questions about this project specifically. I am a better researcher and mathematician because of my interactions with them both.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The annual National Collegiate Athletic Association (NCAA) Division I men's basketball tournament, known as *March Madness*, is one of the most popular sporting events in the United States in terms of viewership and bets placed, with over 97 million viewers from 180 countries watching at least a portion of the tournament in 2018, and billions of dollars gambled every year during the tournament (NCAA, 2018). The cultural and financial weight of the tournament lead to research interest in many fields.

To provide a brief introduction, the tournament is conducted every year after the end of the college basketball season. The tournament consists of 64 teams, 32 conference champions and 32 at-large teams decided by the selection committee. Once the teams are selected, they are divided among four tournament regions and assigned seeds from 1 to 16, with the lower seeds given to better teams. Six rounds of single-elimination games are then played over three consecutive long weekends (late March to mid-April) on neutral courts. A more detailed description of the tournament is given in Chapter 2.

A common way for fans to bet or compete during the tournament is to enter in a *bracket competition* and predict the outcome of all 63 games before any games are played. An estimated of 16.2 million brackets were submitted to the ESPN's men's tournament challenge in 2021 alone (ESPN, 2021), in addition to entries into bracket competitions through a number of online sites designed specifically to assist fans in running their own bracket pools.

At the time of writing, there has not yet been a perfect bracket prediction. In addition to the large number of possible combinations of game results in a bracket,

there are other factors complicating bracket predictions:

1. Game pairings are inherently biased since lower-seeded teams are given an easier path through the tournament (i.e., the first two games for the 1-seed team in each region are against the 16-seed team and the winner of the 8- vs. 9-seed game).

2. The single-elimination format of the tournament introduces high variability, since the elimination of a team expected to perform well results in the loss of all possible points for that team's predicted wins in the bracket.

3. Bracket predictions are locked in before competition begins. As such, individual player injuries or 'hot streaks' cannot be predicted or adjusted for, but meaningfully impact team performance and game results.

One method that can be used to complete the bracket is according to the tournament seeding, i.e., choosing the team with the smallest seed number as the team to advance to the next round. Although a team's seed has been proven to be a strong predictor of victory in games (Smith, 1999), tournament seeding is subjective and potentially biased. By calculating the distribution of the seeds of winning teams by round from 1985 to 2010, Jacobson et al. (2011) show that although the winning team is most frequently a 1-seed, all four teams of the Final 4 very rarely are. For that reason, and to incorporate other relevant game statistics, other methods have been developed that account for a more global picture of a team's performance.

Common team ranking methods used to evaluate teams and to complete a bracket consist of the Pomeroy ratings, the Sagarin ratings, the Massey ratings, and the Ratings Percentage Index (RPI); see Section 5.2 for details. Moreover, Kvam and Sokol (2006) use a logistic regression model to estimate win probabilities through a Markov chain (LRMC) model. Brown and Sokol (2010) suggest an improved LRMC model by using empirical Bayes and ordinary least squares. West (2006, 2008) proposes a rating

method based on ordinal logistic regression and expectation that focuses on calculating the expected number of wins for teams that are selected for the tournament. Koenker and Bassett (2010) propose a quantile regression approach to complete the bracket, while Shen et al. (2015) consider a method based on a binomial generalized linear regression model with Cauchy link. Gupta (2015) proposes a dual-proportion likelihood methodology for competitive bracket advantage. Finally, Ludden et al. (2020) estimate winning probabilities as a power function of the seed number.

A different approach for calculating the probability of a win is to utilize binary quantile regression, where the probability is obtained by *averaging* over multiple quantile levels. This allows for a more comprehensive estimation, as the different quantiles can better describe the entire conditional distribution. Specifically, in this work, we propose the *average binary quantile regression* (ABQR), which uses the dimension reduction method of Christou (2020) for the estimation of the conditional quantiles and then estimates the probability of a win by averaging across a grid of conditional quantile estimates. We note that the dimension reduction component of the methodology is of great interest as in practice, the number of available statistics in sports analysis are only increasing. As such, methods allowing for a large number of predictors in the model are desirable.

To test the performance of the method, we apply it to several tournament years and complete bracket predictions. Our results suggest that ABQR has a good finite sample performance and often outperforms a number of commonly used methods. In fact, the proposed methodology shows a competitive performance on predicting upsets and it outperformed all methods for all scoring schemes during the 2022 tournament. To the best of our knowledge, this is the first work that utilizes multiple quantile levels for the estimation of a win probability in order to complete a bracket.

The rest of the thesis is organized as follows: Chapter 2 provides a detailed description of the tournament and of the different scoring methods. Chapter 3 intro-

duces quantile regression and its applications, while Chapter 4 presents the proposed methodology. Chapter 5 applies the methodology to March Madness prediction and compares results to existing methods. A brief discussion of the results is given in Chapter 6.

CHAPTER 2: TOURNAMENT STRUCTURE AND SCORING

The NCAA Division I men's basketball tournament is made up of 64 teams each year: the champions of the 32 Division I conferences, who are granted automatic bids, and 32 additional teams as selected by the committee. The 64 teams are then divided into four regions and ranked 1 through 16 within each region based on their performance in the regular season, with the best four teams as 1-seeds in their respective regions. For the first round, the teams compete within each region according to their seeding. Specifically, 1 is paired with 16, 2 is paired with 15, and so on, such that the sum of the seeds of each pair is 17. These single-elimination games continue within each region for three more rounds, called the Round 2, the Sweet 16, and the Elite 8. Then, the winners of each region are paired in the next round, the Final 4. Finally, the two surviving teams play the sixth and final championship round, where the winner is declared as the tournament's champion. These six rounds result in a total of 63 games. A visual representation of the tournament is given in Figure 2.1.

Remark. *The NCAA tournament structure has been updated periodically throughout its long history. Most recently, a play-in round was introduced in 2001 and expanded in 2011. Specifically, 68 teams are selected initially, with 60 teams guaranteed tournament spots and the remaining eight teams playing single-elimination games to determine the remaining four tournament teams. These games are called the First 4 and are played prior to the start of the regular tournament. This play-in round is generally not included in bracket scoring and will not be considered here. As such, the proposed methodology focuses only on the 64-team tournament.*

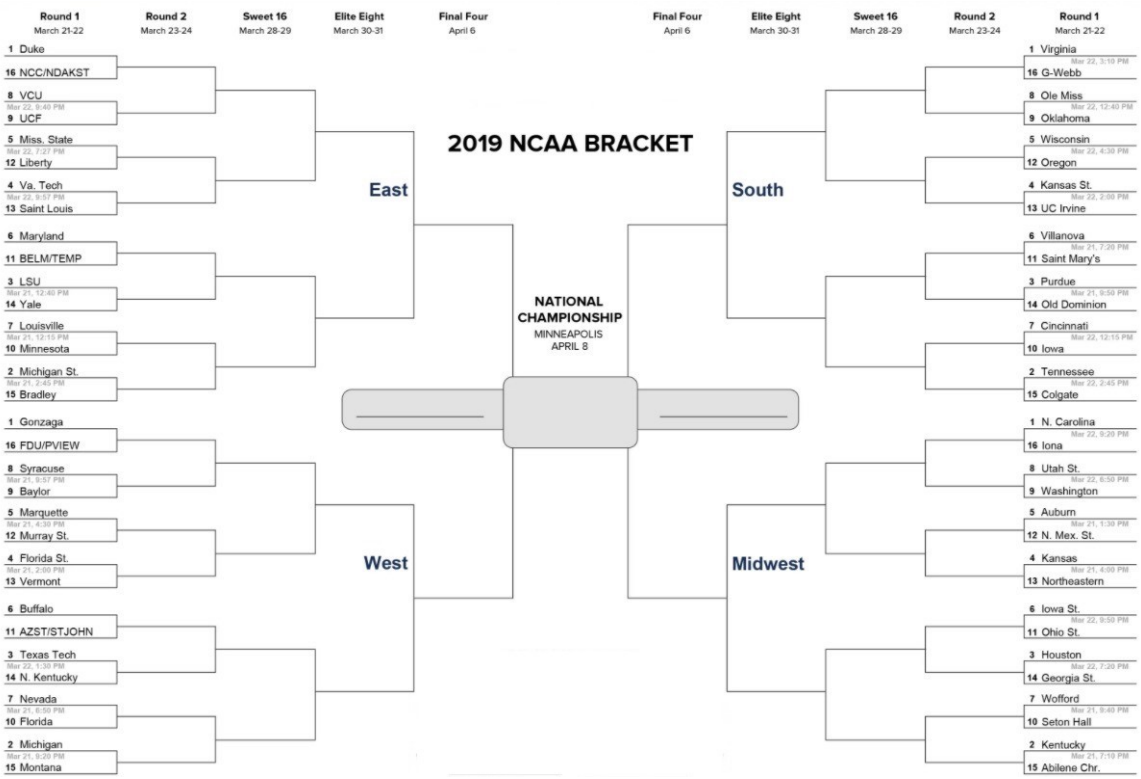

Figure 2.1: Sample Bracket of the 2019 NCAA Tournament.

2.1 Bracket Scoring

Bracket competitions involve choosing the winner of each of the 63 tournament games in full before the start of the first round. One way to evaluate a bracket is by counting how many correct predictions were made; this is called the *single scoring* system. In other words, single scoring awards one point for each correct game prediction, irrespective of the round, resulting in a maximum of 63 points per bracket. In contrast, *double scoring* assigns more weight to later rounds since the cumulative nature of the bracket makes it harder to have correct predictions in advanced rounds. Specifically, double scoring doubles the amount of points awarded for each game predicted by round, i.e., one point for each correct prediction for Round 1, two points for each correct prediction for Round 2, four points for each correct prediction for Round 3, and so on. In this scoring method, each round is worth 32 points, even though the number of games per round is reduced by half. As such, predicting the

correct champion (one game) is worth as many points as the entire first round (32 games). The maximum score under this scoring system is 192 points per bracket. The nature of this scoring system makes it very attractive and it has been used extensively.

Another scoring method that will be used in this thesis is *upset scoring*, which rewards more points for correctly predicting an *upset*, i.e., an outcome where a higher-seeded team defeats a lower-seeded team. There are multiple upset scoring methods in use; one of the most common, and the method used in this work, starts with the double scoring method and multiplies the points earned by the seed of the winning team (Kvam and Sokol, 2006). For example, successfully predicting a win in Sweet 16 by a 6-seed team results in a score of $4 \times 6 = 24$ points, since Round 3 predictions earn four points in double scoring, whereas a similar game prediction involving a 2-seed team results in a score of $4 \times 2 = 8$ points. Upset scoring is less common in bracket competitions, but it can provide useful insights regrading the predictive ability of upsets for different methods. Table 2.1 demonstrates a summary of the three different scoring systems.

Table 2.1: Points awarded for correct predictions per method and round.

	Rd1	Rd2	Sweet16	Elite8	Final4	Championship	Total
Single	1	1	1	1	1	1	63
Double	1	2	4	8	16	32	192
Upset	1×seed	2×seed	4×seed	8×seed	16×seed	32×seed	

CHAPTER 3: QUANTILE REGRESSION

Quantile regression (QR) has been proposed as an alternative to ordinary least squares (OLS) regression, especially in cases where the error term has non-constant variance. Modeling the conditional quantile of a response variable Y given a p-dimensional vector of predictors $\mathbf{X}$ provides a more complete picture of the distribution, where the τth conditional quantile is defined as

$$Q_\tau(Y|\mathbf{x}) = Q_\tau(Y|\mathbf{X}=\mathbf{x}) = \inf\{y : P(Y \le y|\mathbf{X}=\mathbf{x}) \ge \tau\},$$

for $\tau \in (0,1)$.

QR was first introduced by Koenker and Bassett (1978), who consider a linear QR model $Q_\tau(Y|\mathbf{x}) = \alpha_\tau + \boldsymbol{\beta}_\tau^\top \mathbf{x}$, $\alpha_\tau \in R$, $\boldsymbol{\beta}_\tau \in R^p$ and use the representation

$$Q_\tau(Y|\mathbf{x}) = \arg\min_q E\{\rho_\tau(Y-q)|\mathbf{X}=\mathbf{x}\},$$

where $\rho_\tau(u) = \{\tau - I(u<0)\}u$, to define the estimator $(\widehat{\alpha}_\tau, \widehat{\boldsymbol{\beta}}_\tau)$. Specifically, for independent and identically distributed (iid) observations $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$,

$$(\widehat{\alpha}_\tau, \widehat{\boldsymbol{\beta}}_\tau) = \arg\min_{(a_\tau, \mathbf{b}_\tau)} \sum_{i=1}^n \rho_\tau(Y_i - a_\tau - \mathbf{b}_\tau^\top \mathbf{X}_i).$$

Then, $\widehat{\alpha}_\tau + \widehat{\boldsymbol{\beta}}_\tau^\top \mathbf{x}$ gives the estimator of the τth conditional quantile under the linear QR model.

As the linearity assumption is quite strict, several authors considered the completely flexible nonparametric estimation of the conditional quantiles; see Chaudhuri (1991), Yu and Jones (1998), and Guerre and Sabbah (2012), among others. How-

ever, nonparametric estimation techniques suffer from the well-known 'curse of dimensionality' problem and for that reason many researchers turn their attention to semiparametric models.

A semiparametric model that received particular attention is the single index quantile regression (SIQR) model that defines

$$Q_\tau(Y|\mathbf{x}) = g_\tau(\boldsymbol{\beta}_\tau^\top \mathbf{x}),$$

for $g_\tau(\cdot) : R \to R$ an unknown univariate link function, called the nonparametric component, and $\boldsymbol{\beta}_\tau \in R^p$ is a fixed, but unknown, vector of parameters, called the parametric component. This model assumes that we can replace the p-dimensional predictor vector $\mathbf{X}$ with the one-dimensional predictor vector $\boldsymbol{\beta}_\tau^\top \mathbf{X}$ without losing any important information necessary for the estimation of the conditional quantile. This model has been considered by serveral authors, including Wu et al. (2010), Kong and Xia (2012), and Christou and Akritas (2016). A generalization of the above model is the multi-index quantile regression (MIQR) model that assumes that

$$Q_\tau(Y|\mathbf{x}) = g_\tau(\mathbf{B}_\tau^\top \mathbf{x}),$$

where $\mathbf{B}_\tau$ is a $p \times d_\tau$ matrix of unknown coefficients, and $1 \leq d_\tau \leq p$. This model also received special attention by Luo et al. (2014) and Christou (2020).

For a further reading of the applications of QR across multiple fields of study, see the works of Yu et al. (2003), who present practical applications in medicine, economics, survival analysis, and detecting heteroskedasticity, and Leeds (2014), who proposes multiple uses for QR in sports economics. Moreover, QR has been used to consider covariates affecting athlete salary; see Burnett and Van Scyoc (2013, 2015), Deutscher and Büschemann (2014), and Vincent and Eastman (2009).

CHAPTER 4: PROPOSED METHODOLOGY

4.1 Binary Quantile Regression Approach

Let Y be a univariate binary response and $\mathbf{X}$ be a p-dimensional vector of predictors. The goal is to estimate the probability $P(Y = 1|\mathbf{X} = \mathbf{x})$ using observed data. A common and popular technique is the logistic regression model. However, an attractive alternative, as proposed by Kordas (2006), treats the probability as an average of multiple conditional quantiles. This allows for a more complete picture of the conditional probability. Below we outline the methodology.

Let Y^* be a scalar latent continuous variable, where Y is its observed binary indicator. In this work, we assume the model

$$\begin{aligned} Q_\tau(Y^*|\mathbf{x}) &= g_\tau(\mathbf{B}_\tau^\top \mathbf{x}), \\ Y &= I\{Y^* \geq 0\}, \end{aligned} \tag{4.1}$$

where $\mathbf{B}_\tau$ is a $p \times d_\tau$ matrix of unknown parameters, $d_\tau \leq p$, and $g_\tau(\cdot) : R^{d_\tau} \to R$ is a d_τ-dimensional link function. This model allows for nonparametric flexibility, while at the same time reduces the dimension of the predictor variables. Although there are many existing methods for fitting the above model, we will use the methodology of Christou (2020), as it has a competitive finite sample performance. For a smoother reading of this section, we present an explanation of this method in Section 4.2.

Since Y^* is a latent variable, we need to find a way to express the model in terms of Y. The equivariance property of quantile functions (Powell 1984, 1986) states that

$$f\{Q_\tau(Y^*|\mathbf{x})\} = Q_\tau\{f(Y^*)|\mathbf{x}\},$$

for any real, monotone, increasing function $f(\cdot)$. This implies that

$$I\{Q_\tau(Y^*|\mathbf{x}) \geq 0\} = Q_\tau\{I(Y^* \geq 0)|\mathbf{x}\} = Q_\tau(Y|\mathbf{x}). \tag{4.2}$$

Therefore, relationships (4.1) and (4.2) imply that

$$Q_\tau(Y|\mathbf{x}) = I\{g_\tau(\mathbf{B}_\tau^\top \mathbf{x}) \geq 0\} = \widetilde{g}_\tau(\mathbf{B}_\tau^\top \mathbf{x}), \tag{4.3}$$

for some d_τ-dimensional function $\widetilde{g}_\tau(\cdot)$.

Kordas (2006) uses the fact that

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \int_0^1 I\{g_\tau(\mathbf{B}_\tau^\top \mathbf{x}) \geq 0\} d\tau = \int_0^1 \widetilde{g}_\tau(\mathbf{B}_\tau^\top \mathbf{x}) d\tau$$

to express the probability as an integral over the quantiles levels. This idea leads to the following estimation procedure. Let $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ denote the set of n observations. First, we use existing dimension reduction techniques to estimate the column vectors of the matrix $\mathbf{B}_\tau$, denoted by $\widehat{\mathbf{B}}_\tau$, and form the new sufficient predictors $\widehat{\mathbf{B}}_\tau^\top \mathbf{X}_i$, $i = 1, \ldots, n$. In this work, we use the τth central quantile subspace (τ-CQS) of Christou (2020), as explained in more detail in Section 4.2. Next, we use nonparametric techniques to estimate $Q_\tau(Y|\widehat{\mathbf{B}}_\tau^\top \mathbf{X}_i) = \widetilde{g}_\tau(\widehat{\mathbf{B}}_\tau^\top \mathbf{X}_i)$. For that, we use the local linear conditional quantile estimation method of Yu and Jones (1998), as it tends to work well in practice. Specifically, we take $\widehat{Q}_\tau(Y|\widehat{\mathbf{B}}_\tau^\top \mathbf{X}_i) = \widehat{q}_\tau(\mathbf{X}_i)$, where

$$\begin{aligned}(\widehat{q}_\tau(\mathbf{X}_i), \widehat{\mathbf{s}}_\tau(\mathbf{X}_i)) = \arg\min_{(q_\tau, \mathbf{s}_\tau)} \sum_{k=1}^n \rho_\tau \left\{ Y_k - q_\tau - \mathbf{s}_\tau^\top \widehat{\mathbf{B}}_\tau^\top (\mathbf{X}_k - \mathbf{X}_i) \right\} \\ \times K\left\{ \frac{\widehat{\mathbf{B}}_\tau^\top (\mathbf{X}_k - \mathbf{X}_i)}{h} \right\}.\end{aligned} \tag{4.4}$$

Here, $K(\cdot)$ is a d_τ-dimensional kernel function and $h > 0$ is a bandwidth. In this work, we use a Gaussian kernel and choose the bandwidth using the optimal one

of mean regression local estimation, i.e., $h = sd(y)n^{-1/5}$, where $sd(y)$ denotes the sample standard deviation of the n observations; see Fan and Gijbels (1995) and Ruppert et al. (1995). Finally, we estimate the probability on a grid of quantile levels. Specifically,

$$\widehat{P}(Y = 1|\mathbf{X} = \mathbf{X}_i) = \sum_{k=1}^{K} \widehat{Q}_{\tau_k}(Y|\widehat{\mathbf{B}}_\tau^\top \mathbf{X}_i)(\tau_k - \tau_{k-1}),$$

where $\tau_1 < \cdots < \tau_K$ are the quantile levels, and K is the number of grid points. For simplicity, we assume evenly spaced quantiles, i.e., $\tau_k = k/K$, for $k = 1, 2, \ldots, K$. This leads to the estimator

$$\widehat{P}(Y = 1|\mathbf{X} = \mathbf{X}_i) = \frac{1}{K} \sum_{k=1}^{K} \widehat{Q}_{\tau_k}(Y|\widehat{\mathbf{B}}_\tau^\top \mathbf{X}_i).$$

4.2 Estimation of the τth Central Quantile Subspace

We now discuss how to estimate $\mathbf{B}_\tau$ using the methodology proposed by Christou (2020). Relationship (4.3) implies that $\mathbf{B}_\tau^\top \mathbf{X}$ contains all the information we need to know about the conditional quantile function and therefore, the goal is to estimate the matrix $\mathbf{B}_\tau$ in order to form the new reduced predictors. However, since the function $\widetilde{g}_\tau$ is nonparametric, $\mathbf{B}_\tau$ is identifiable up to a constant. That is, the interest is on the space spanned by $\mathbf{B}_\tau$, called the τth central quantile subspace (τ-CQS) and denoted by $\mathcal{S}_{Q_\tau(Y|\mathbf{X})}$.

This methodology (Christou, 2020) is based on two important statements:

(a) $\boldsymbol{\beta}_\tau \in \mathcal{S}_{Q_\tau(Y|\mathbf{X})}$, where $(\alpha_\tau, \boldsymbol{\beta}_\tau) = \arg\min_{(a_\tau, \mathbf{b}_\tau)} E\{Q_\tau(Y|\mathbf{X}) - a_\tau - \mathbf{b}_\tau^\top \mathbf{X}\}^2$.

(b) $E\{Q_\tau(Y|U_\tau)\mathbf{X}\} \in \mathcal{S}_{Q_\tau(Y|\mathbf{X})}$, where U_τ is a measurable function of $\mathbf{B}_\tau^\top \mathbf{X}$, provided that $Q_\tau(Y|U_\tau)\mathbf{X}$ is integrable.

Part (a) implies that the ordinary least squares (OLS) vector $\boldsymbol{\beta}_\tau$, resulting from

regressing $Q_\tau(Y|\mathbf{X})$ on $\mathbf{X}$, belongs to $\mathcal{S}_{Q_\tau(Y|\mathbf{X})}$. This provides a single vector in the τ-CQS. If the dimension d_τ of the τ-CQS is one, then one can stop and report $\boldsymbol{\beta}_\tau$ as the vector that will define the new sufficient predictor $\boldsymbol{\beta}_\tau^\top \mathbf{X}$. However, if the dimension d_τ is greater than one, then the single OLS vector is insufficient. Part (b) proposes a method to produce more vectors in $\mathcal{S}_{Q_\tau(Y|\mathbf{X})}$. Specifically, setting $\boldsymbol{\beta}_{\tau,0} = \boldsymbol{\beta}_\tau$, we define, for $j = 1, \ldots, p-1$ and a function $u_\tau(\cdot)$,

$$\boldsymbol{\beta}_{\tau,j} = E[Q_\tau\{Y|u_\tau(\boldsymbol{\beta}_{\tau,j-1}^\top \mathbf{X})\}\mathbf{X}] \in \mathcal{S}_{Q_\tau(Y|\mathbf{X})}.$$

Christou (2020) uses the identity function $u_\tau(t) = t$. Finally, for a $p \times p$ matrix $\mathbf{V}_\tau = (\boldsymbol{\beta}_{\tau,0}, \ldots, \boldsymbol{\beta}_{\tau,p-1})$, we perform an eigenvalue decomposition of $\mathbf{V}_\tau \mathbf{V}_\tau^\top$ and choose the d_τ eigenvectors corresponding to the d_τ nonzero eigenvalues.

The estimation procedure can be summarized as follows. First, estimate $Q_\tau(Y|\mathbf{X}_i)$ using a nonparametric technique. Note that nonparametric techniques face the so-called 'curse of dimensionality' problem when the dimension of the predictor vector $\mathbf{X}$ is large. For that reason, Christou (2020) proposes performing an initial dimension reduction technique and replace $\mathbf{X}$ with $\mathbf{A}^\top \mathbf{X}$, for $\mathbf{A}$ a $p \times d$ matrix, $d \leq p$. For more information, see Li (1991). Then, take $\widehat{Q}_\tau(Y|\mathbf{X}_i) = \widehat{q}_\tau(\mathbf{X}_i)$, where $\widehat{q}_\tau(\mathbf{X}_i)$ is given by (4.4), except that we replace $\widehat{\mathbf{B}}_\tau$ with $\widehat{\mathbf{A}}$. Next, estimate $\boldsymbol{\beta}_\tau$ by

$$(\widehat{a}_\tau, \widehat{\boldsymbol{\beta}}_\tau) = \arg\min_{(a_\tau, \mathbf{b}_\tau)} \sum_{i=1}^{n} \{\widehat{Q}_\tau(Y|\mathbf{X}_i) - a_\tau - \mathbf{b}_\tau^\top \mathbf{X}_i\}^2.$$

Following, set $\widehat{\boldsymbol{\beta}}_{\tau,0} = \widehat{\boldsymbol{\beta}}_\tau$, and for $j = 1, \ldots, p-1$, define

$$\widehat{\boldsymbol{\beta}}_{\tau,j} = n^{-1} \sum_{i=1}^{n} \widehat{Q}_\tau(Y|\widehat{\boldsymbol{\beta}}_{\tau,j-1}^\top \mathbf{X}_i)\mathbf{X}_i,$$

where $\widehat{Q}_\tau(Y|\widehat{\boldsymbol{\beta}}_{\tau,j-1}^\top \mathbf{X}_i)$ is obtained using (4.4) but with $\widehat{\mathbf{B}}_\tau$ replaced by $\widehat{\boldsymbol{\beta}}_{\tau,j-1}$. Finally, define $\widehat{\mathbf{V}}_\tau = (\widehat{\boldsymbol{\beta}}_{\tau,0}, \ldots, \widehat{\boldsymbol{\beta}}_{\tau,p-1})$ and choose the eigenvectors $\widehat{\mathbf{v}}_{\tau,k}$, $k = 1, \ldots, d_\tau$,

corresponding to the d_τ largest eigenvalues of $\widehat{\mathbf{V}}_\tau \widehat{\mathbf{V}}_\tau^\top$. Then, $\widehat{\mathbf{B}}_\tau = (\widehat{\mathbf{v}}_{\tau,1}, \ldots, \widehat{\mathbf{v}}_{\tau,d_\tau})$ is an estimated basis matrix for $\mathcal{S}_{Q_\tau(Y|\mathbf{X})}$. That means that $\widehat{\mathbf{B}}_\tau$ is an estimate of $\mathbf{B}_\tau$ up to a multiplier, which is taken into account by the nonparametric link function $\widetilde{g}_\tau$.

Remark. *In practice, the true dimension d_τ is unknown and needs to be estimated. In this work, we estimate d_τ using the modified Bayesian information criterion (BIC) of Zhu et al. (2010).*

4.3 Use in Bracket Prediction

We now discuss how to apply the proposed methodology in order to fill out a bracket. First, we note that a bracket consists of six rounds; therefore, the method is repeated for each round by taking into account which teams were predicted to advance to that round. For each tournament team we have six different binary responses, $Y^{(1)}, \ldots, Y^{(6)}$, indicating whether a team won or lost on the lth round, $l = 1, \ldots, 6$, and a vector of predictors $\mathbf{X}$ characterizing the team's performance during the regular season. There are 64 teams at the start of each tournament, so the full data set consists of $64 \times T$ observations, where T denotes the number of years we use as historical data.

For the first round, we use all $64 \times T$ observations as our training set in order to fit the model; the response variable is $Y^{(1)}$. For the second round, we use $32 \times T$ observations, consisting of all the observations from the teams that made it to the second round; the response variable is $Y^{(2)}$. We continue similarly for the rest of the rounds. For example, for the sixth and final round, we use $2 \times T$ observations, consisting of all the observations from the two final teams that made it to the championship during the last T years; the response variable is $Y^{(6)}$. For each round, we fit the model using historical data and then use the current year's observations to predict the probability of each team winning. We now present the algorithm.

Algorithm: For what follows, $\mathbf{X}^*$ denotes a team's regular season performance for

the year under consideration.

1. Let $\{Y_i^{(1)}, \mathbf{X}_i\}_{i=1}^{64\times T}$ denote the historical data of the first round for T years.

2. Given $\tau_k = k/K$, repeat for $k = 1, \ldots, K$,

   (a) use the method of Christou (2020), described in Section 4.2, to estimate $\mathbf{B}_{\tau_k}$ and form the new predictors $\widehat{\mathbf{B}}_{\tau_k}^\top \mathbf{X}_i$.

   (b) estimate $Q_\tau(Y^{(1)}|\widehat{\mathbf{B}}_{\tau_k}^\top \mathbf{X}^*)$ using (4.4) for each team playing in that round; denote the estimate with $\widehat{Q}_\tau(Y^{(1)}|\widehat{\mathbf{B}}_{\tau_k}^\top \mathbf{X}^*)$.

3. For each team that is playing in that round, estimate the probability of winning using

$$\frac{1}{K}\sum_{k=1}^{K} \widehat{Q}_{\tau_k}(Y^{(1)}|\widehat{\mathbf{B}}_{\tau}^\top \mathbf{X}^*).$$

4. Pair all the teams according to the bracket seeding in Round 1 and choose the team with the highest probability of each pair to advance.

5. Update $\mathbf{X}^*$ as the performance of teams predicted to advance to Round 2.

6. Given round l, repeat for $l = 2, \ldots, 5$,

   (a) let $\{Y_i^{(l)}, \mathbf{X}_i\}_{i=1}^{64\times T/2^{l-1}}$ denote the historical data of all the teams that advanced in round l for T years.

   (b) Repeat Steps 2 - 3.

   (c) Pair all the teams according to the predictions of the previous round and choose the teams to advance according to the highest probability.

   (d) Update $\mathbf{X}^*$ as any team's performance that is predicted to advance to round $l + 1$.

7. For the final round, let $\{Y_i^{(6)}, \mathbf{X}_i\}_{i=1}^{2\times T}$ denote the historical data of the championship games for T years.

8. Repeat Steps 2 - 3.

9. Declare as champion the team with the higher probability.

Remarks.

(a) *Step 2 (a) is performed using the* `cqs` *function of the* `quantdr` *package in* `R` *and Step 2 (b) is performed using the* `llqr` *function of the same package.*

(b) For this work, we use $K = 20$ *to create the grid of quantile levels. We note that we also tried different values of K, but there was no major difference in the results.*

CHAPTER 5: TOURNAMENT APPLICATION AND RESULTS

5.1 Computational Remarks

In this section, the proposed methodology is applied to several years' March Madness tournaments and its performance is compared with other commonly used methods. The data consist of observations from the 64 tournament teams for the years 2002-2022; there was no tournament held in 2020, so 2020's regular season statistics were excluded. As was mentioned in Section 4.3, the response variable differs for each round and consists of a binary variable indicating whether a team won or lost during that round, assuming that the team has advanced to that round. The predictor variables consist of 14 game statistics averaged across each season and the seed number. These averaged variables include three-pointers per game, field goals per game, free throw attempts per game, free throws per 100 possessions, offensive rebound percentage, offensive rebounds per game, defensive rebound percentage, defensive rebounds per game, assists per game, fouls per game, scoring margin, assist-to-turnover ratio, offensive efficiency, and defensive efficiency. Region, seeding, and tournament results are obtained from htpps://www.ncaa.com. The 14 regular season statistics are obtained from https://www.teamrankings.com.

To demonstrate the performance of the proposed methodology and compare it with existing methods, bracket results were predicted for five seasons: 2016, 2017, 2018, 2019, and 2022. Note that there was no tournament played in 2020 due to COVID-19 restrictions. While a tournament was played in 2021, the regular season statistics were not a fair comparison for the field of 64 teams as different COVID-19 protocol by conference led to unusually high variance in number of quarantined players and games played per team.

To predict the results for the 201x tournament, we used data from years 2002 up to the year before the current tournament. We note that we also tried our methodology using a moving window, but we found that the results under performed when compared to the full training set; for that reason, all previous years were used in the training set.

Remark. *A conscious decision was made to consider only publicly available game statistics as predictor variables. If a method relies on a metric calculated by a third party, such as Sagarin's Schedule ranking or NCAA Evaluation Tool (NET), the method cannot be applied for evaluating seasons outside the available range of the third-party metric. In contrast, our method does not rely on any such metric and is easily reproducible for past and future tournament years with publicly available team data.*

5.2 Methods Considered for Comparison

The methods used for comparison brackets are as follows:

1. The *seed* of a team, assigned by the selection committee, can be used to fill out a bracket. Specifically, for a single game, the team with the lowest seed will be selected to advance to the following round. Team seed is listed on the official tournament bracket released by the NCAA.

2. *Pomeroy's College Basketball Ratings* are based on the Pythagorean winning percentage formula given in Kubatko et al. (2007), which takes into account the adjusted offensive efficiency (AdjO) and the adjusted defensive efficiency (AdjD) of a team. In a single game, the team with the lowest Pomeroy rating will be selected to advance to the following round. The Pomeroy ratings are available in https://kenpom.com.

3. *Jeff Sagarin's College Basketball Ratings* have been available for use in bracket completion since 1985. Although the details of the method are not publicly

available, the ratings are based on two characteristics: Sagarin's personal modification of the Elo chess rating system (Elo, 1978), and a rating method developed by Sagarin known as the 'Pure Points' method (West, 2006). In a single game, the team with the lowest Sagarin rating will be selected to advance. The Sagarin ratings can be found in http://sagarin.com/sports/cbsend.htm.

4. The *Logistic Regression/Markov Chain (LRMC)* method, proposed by Kvam and Sokol (2006), uses a logistic regression model on basic scoreboard data in order to estimate the probability that one team is better than another. These probabilities are used as transition probabilities in a Markov Chain model in order to determine a ranking system, where, in a single game, the team with the lowest rating will be selected to advance. We note that Brown and Sokol (2010) modified this methodology by replacing the logistic regression part with an empirical Bayes model. However, in this work, we compare our method with the classic LRMC. Both classic and Bayesian LRMC ratings can be found at https://www2.isye.gatech.edu/ jsokol/lrmc/.

5. The *Massey ratings* are designed to measure past performance of teams and are calculated by applying a Bayesian win-loss correction to ratings based on game score, location, and date. In a single game, the team with the lowest Massey rating will be selected to advance. The Massey ratings can be found in https://masseyratings.com/cb/ncaa-d1/ratings.

6. The *Ratings Percentage Index (RPI)* is a rating method based on the team's performance and strength of schedule. Specifically,

$$RPI = 0.25WP + 0.5OWP + 0.25OOWP,$$

where WP, OWP and $OOWP$ denote the winning percentage, the opponents'

winning percentage, and the opponents' opponents' winning percentage, respectively. In a single game, the team with the highest RPI will be predicted to win the game. The RPI can be found at https://masseyratings.com/cb/compare.htm.

Remark. *From 1981–2018, RPI was used by the selection committee as a component of the tournament seeding process. However, following the 2018-2019 season, RPI was discontinued in favor of a new evaluation metric (NCAA Evaluation Tool–NET). We have chosen to report the results for bracket prediction using RPI for 2016-2018 because RPI was regularly used as a benchmark in literature prior to 2019. This allows the proposed method's scores to be easily compared with those in earlier works.*

5.3 Results

The results for predicting the outcome of the tournament for the years tested are given in Tables 5.1-5.3. For each method, the total number of points throughout the years is demonstrated with the largest value each year underlined. In case of ties, all relevant values are underlined. Note that the total for RPI only corresponds to years 2016-2018, since RPI was not available from 2019 and afterwards.

According to the single scoring system, all methods seem to predict between 32 and 46 correct games, with the Sagarin method having a larger overall prediction by a small margin. The double scoring system demonstrates how each method correctly predicts games progressively in later rounds. We observe that the proposed methodology (ABQR), Pomeroy, and Sagarin seem to have the best overall performance. Specifically, the top score in each year comes from one of the three methods mentioned: ABQR outperforms all existing methods for 2016 and 2022, Sagarin performs the best for 2017 and 2018, and Pomeroy beats all methods with a large difference for 2019. Over all years combined, the proposed methodology has the largest total with 490 points. According to the upset scoring results, ABQR, Pomeroy, and LRMC have the best overall performance, with the proposed methodology having the largest total of 1128 points.

Table 5.1: Number of points accumulated for each method and year according to the single scoring system.

Year	ABQR	Seed	Pomeroy	Sagarin	LRMC	Massey	RPI
2016	38	37	39	**40**	**40**	39	38
2017	45	44	44	**46**	44	43	39
2018	39	36	38	39	**40**	39	38
2019	41	41	**44**	43	42	41	N/A
2022	**38**	32	33	35	35	35	N/A
Total	201	190	198	**203**	201	197	115

Table 5.2: Number of points accumulated for each method and year according to the double scoring system.

Year	ABQR	Seed	Pomeroy	Sagarin	LRMC	Massey	RPI
2016	**115**	87	79	82	93	88	73
2017	102	82	110	**113**	88	90	61
2018	88	81	78	**111**	110	79	84
2019	97	92	**127**	94	93	88	N/A
2022	**88**	62	45	57	48	54	N/A
Total	**490**	404	439	457	432	399	218

Table 5.3: Number of points accumulated for each method and year according to the upset scoring system.

Year	ABQR	Seed	Pomeroy	Sagarin	LRMC	Massey	RPI
2016	**269**	169	195	221	230	193	157
2017	**246**	206	232	244	244	214	168
2018	203	172	196	231	**240**	210	198
2019	203	182	**233**	204	195	191	N/A
2022	**207**	161	146	160	163	161	N/A
Total	**1128**	890	1017	1046	1069	971	523

The results suggest that the proposed methodology performs very well for predicting correct results when weight is given to later rounds and in predicting upsets.

CHAPTER 6: DISCUSSION

In this work we proposed a new method for predicting the results of the NCAA basketball tournament. The method estimates the conditional quantiles of the observed binary response given a set of predictor variables that characterize the team's performance throughout the season. Then, the winning probability of each team is calculated as the average of multiple conditional quantiles. The real-world data analysis shows that ABQR has a competitive performance and often outperforms other commonly-used methods. Moreover, the proposed methodology relies on a simple-to-implement computational algorithm that uses only publicly available team data. To the best of our knowledge, this is the first work that approaches a conditional probability by averaging over a grid of estimated conditional quantiles and we hope that this work will stimulate further interest in quantile regression both within and out of the context of sports analytics.

The promising results of the proposed method invite other extensions or areas of study. Currently, there is no consideration for strength of schedule, conference affiliation, or game strategy in the model; future work could consider these or other relevant predictors. The proposed model uses team performance data from the entire season, while future models may consider incorporating a momentum component for teams that seem to hit a 'hot streak' and improve dramatically as the tournament begins. Additionally, the proposed method and algorithm can be applied to other single-elimination tournaments, such as the NCAA Division I women's basketball tournament, the College Football Playoff, or the Football Association Cup.

Moreover, defining a different 'success' for the binary variable Y opens possibilities beyond March Madness brackets. For example, calculating a team's probability of

covering the spread may be of interest in the realm of sports gambling and would give a better picture of team success than simply winning or losing due to the inherently biased nature of seeding.

REFERENCES

[1] Brown, M. and J. Sokol. 2010. "An Improved LRMC Method for NCAA Basketball Prediction." *Journal of Quantitative Analysis in Sports*, 6(3):1–23.

[2] Burnett, N. and L. Van Scyoc. 2013. "Compensation Discrimination for Wide Receivers: Applying Quantile Regression to the National Football League." *Journal of Reviews on Global Economics*, 2:433–441.

[3] ———. 2015. "Compensation Discrimination for Defensive Players: Applying Quantile Regression to the National Football League Market for Linebackers and Offensive Linemen." *Journal of Sports Economics*, 16(4):375–389.

[4] Chaudhuri, P. 1991. "Nonparametric estimates of regression quantiles and their local Bahadur representation." *The Annals of Statistics*, 19(2):760–777.

[5] Christou, E. 2020. Central quantile subspace. *Statistics and Computing*, 30:677–695.

[6] Christou, E. and M. Akritas. 2016. "Single index quantile regression for heteroscedastic data." *Journal of Multivariate Analysis*, 150:169–182.

[7] Deutscher, C. and A. Büschemann. 2014. "Does Performance Consistency Pay Off for Players? Evidence from Bundesliga." *Journal of Sports Economics*, 17(1):27–43.

[8] Elo, A.E. 1978. *The Rating of Chessplayers, Past and Present*. New York: Arco.

[9] ESPN Press Room. 2021. "2021 Tournament Challenge: 16.2 Million Brackets Created." Last modified March 19, 2021. https://espnpressroom.com/us/press-releases/2021/03/2021-tournament-challenge-16-million-brackets-created/

[10] Fan, J. and I. Gijbels. 1995. "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation." *Journal of the Royal Statistical Society, Series B*, 57:371–394.

[11] Guerre, E. and C. Sabbah. 2012. "Uniform bias study and Bahadur representation for local polynomial estimators of conditional quantile function." *Econometric Theory*, 28(1):87–129.

[12] Gupta, A.A. 2015. "A new approach to bracket prediction in the NCAA Men's Basketball Tournament based on a dual-proportion likelihood." *Journal of Quantitative Analysis in Sports*, 11(1):53–67.

[13] Jacobson, S., A. Nikolaev, D. King, and A. Lee. 2011. "Seed distributions for the NCAA men's basketball tournament." *Omega*, 39:719–724.

[14] Koenker, R. and G. Bassett. 1978. "Regression quantiles." *Econometrica*, 46:33–50.

[15] ———. 2010. "March Madness, Quantile Regression Bracketology, and the Hayek Hypothesis." *Journal of Business & Economic Statistics*, 28(1):26–35.

[16] Kong, E. and Y. Xia. 2012. "A single-index quantile regression model and its estimation." *Econometric Theory*, 28(4):730–768.

[17] Kordas, G. 2006. "Smoothed Binary Regression Quantiles." *Journal of Applied Econometrics*, 21(3):387–407.

[18] Kubatko, J., D. Oliver., K. Pelton, and D.T. Rosenbaum. 2007. "A Starting Point for Analyzing Basketball Statistics." *Journal of Quantitative Analysis in Sports*, 3(3):1–18.

[19] Kvam, P. and J. Sokol. 2006. "A logistic regression/Markov chain model for NCAA basketball." *Naval Research Logistics*, 53:788–803.

[20] Leeds, M. 2014. "Quantile Regression for Sports Economics." *International Journal of Sport Finance*, 9:346–359.

[21] Li, K.C. 1991. "Sliced inverse regression for dimension reduction." *Journal of the American Statistical Association*, 86(414):316–327.

[22] Ludden, I., A. Khatibi, D. King, and S. Jacobson. 2020. "Models for generating NCAA men's basketball tournament bracket pools." *Journal of Quantitative Analysis in Sports*, 16(1):1–15.

[23] Luo, W., B. Li, and X. Yin. 2014. "On efficient dimension reduction with respect to a statistical functional of interest." *The Annals of Statistics*, 42(1):382–412.

[24] NCAA. 2018. "2018 NCAA tournament viewership, attendance numbers." Last modified April 13, 2018. https://www.ncaa.com/news/basketball-men/article/2018-04-13/2018-ncaa-tournament-and-final-four-viewership-attendance

[25] Powell, J. 1984. "Least absolute deviations estimation for the censored regression model." *Journal of Econometrics*, 25:303–325.

[26] ———. 1986. "Censored regression quantiles." *Journal of Econometrics*, 32:143–155.

[27] Ruppert, D., S.J. Sheather, and M.P. Wand. 1995. "An effective bandwidth selector for local least squares regression." *Journal of the American Statistical Association*, 90:1257–1270.

[28] Shen, G., S. Hua, X. Zhang, Y. Mu, and R. Magel. (2015). "Predicting Results of March Madness Using the Probability Self-Consistent Method." *International Journal of Sports Science*, 5(4):139–144.

[29] Smith, T. and N.C. Schwertman. 1999. "Can the NCAA basketball tournament seeding be used to predict margin of victory?" *The American Statistician*, 53:94–98.

[30] Vincent, C. and B. Eastman. 2009. "Determinents of Pay in the NHL: A Quantile Regression Approach." *Journal of Sports Economics*, 10(3): 256–277.

[31] West, B.T. 2006. "A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament." *Journal of Quantitative Analysis in Sports*, 2(3).

[32] ———. 2008. "A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament: Updated Results from 2007." *Journal of Quantitative Analysis in Sports*, 4(2).

[33] Wu, T.Z., K. Yu, and Y. Yu. 2010. "Single-index quantile regression." *Journal of Multivariate Analysis*, 101(7):1607–1621.

[34] Yu, K., and Jones, M.C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93(441):228–237.

[35] Yu, K., L. Zudi, and J. Sander. 2003. "Quantile regression: applications and current research areas." *The Statistician*, 52(3):331–350.

[36] Zhu, L.P., L.X. Zhu, and Z.H. Feng. 2010. "Dimension reduction in regression through cumulative slicing estimation." *Journal of the American Statistical Association*, 105(492):1455–1466.