BRACKETING NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Xiao Zhang

In Partial Fulfillment
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

May 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Bracketing NCAA Men's Division I Basketball Tournament

By

Xiao Zhang

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Gang Shen
Chair

Dr. Rhonda Magel

Dr. Seung Won Hyun

Dr. Sean Sather-Wagstaff

Approved:

07/03/2013
Date

Rhonda Magel
Department Chair

ABSTRACT

This paper presents a new bracketing method for all 63 games in the NCAA Division 1 basketball tournament. This method, based on the logistic conditional probability models, is self-consistent in terms of constructing winning probabilities of each game. Empirical results show that this method outperforms the ordinal logistic regression and expectation method with restriction(Restricted OLRE model) proposed by West (2006).

ACKNOWLEDGMENTS

It is with immense gratitude that I acknowledge the support and help of my advisor, Dr. Gang Shen. Thank you for giving me continuous guidance and support throughout this whole research process. I would like to thank all my committee members, Chair Rhonda Magel, Dr. Seung Won Hyun, and Dr. Sean Sather-wagstaff, as well as Mr. Curt Doetkott. At last, I would like to thank all professors and staff from the Department of Statistics.

I owe my deepest gratitude to my dear mother, Baifen Yan, who encourages me all the time and gives me all moral support that I need. I also want to thank my father, sister, girlfriend, and friends who are always by my side.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Research Objective

Among all the national sport games, the annual National Collegiate Athletic Association (NCAA) Men's Division I Basketball Tournament, known as March Madness or the Big Dance, might be the most betted on sporting event in the United States. The tournament is played every spring from March to early April in all neutral venues. This work exclusively focuses on bracketing the NCAA Men's Division I Basketball Tournament by using two major rating systems and two statistical models.

1.2. The History of NCAA Men's Division I Basketball Tournament

The format of the NCAA Men's Division I Basketball Tournament has been changed a couple of times since 1939. The NCAA tournament has expanded a few times throughout the whole history from 8 teams to 68 teams, but the format was the same from 2001 to 2010.

During this period, a selection committee selected a total of 65 teams each year from the 4 regions-namely, Midwest, West, East, and South-for the tournament. In the tournament, there are three regions with 16 teams per region and one region with 17 teams each year. In the region with 17 teams, an opening round game (commonly known as the play-in game) pairs two teams with a 16 seed, and the two teams fight to become the single 16 in that region. Then all 64 teams are seeded from 1 to 16 within their regions. Therefore, while making the bracket, all information about teams' seeds are available.

Since 2011, the tournament has been expanded to 68 teams playing during March and April. Eight of 68 teams first play four games, which are called the First4. After the first four games, 64 teams are finally determined and seeded with a single seed number from 1 to 16 within their regions.

1.3. The Playing Rule and Structure

Then the tournament uses the *single elimination rule* to determine the winner of each game for entering the next round. In terms of bracketing the result of NCAA Men's Division I

Basketball Tournament, our work only considers the last six rounds of the tournament (excluding the First4), namely, Round64 (Rd64), Round32 (Rd32), Sweet16, Elite8, Final4 and the Championship, which have a total of 63 games. Brown and Sokol (2010) refer to it as a bracket. From Rd64, all seasons have the same number of participating teams in each round. Respectively, 64 teams play 32 games in Rd64, 32 teams play 16 games in Rd32, 16 teams play 8 games in Sweet16, 8 teams play 4 games in Elite8, 4 teams play 2 games in Final4, and 2 teams fight for the Championship. Figure 1 shows the diagram of an NCAA regional basketball tournament, and Figure 2 shows the current NCAA Men's Division I Basketball tournament structure in recent seasons.

All 64 teams are seeded by selection committee from 1 through 16 in each of the four regions, from best to worst, and strongest to weakest, according to their relative strength. The committee aims to keep the average strength of each team approximately equal to each other across regions. Breiter and Carlin (1997) explain how a bracket works. In Rd64, each game is paired off so that the seed numbers add up to 17. When the favored winner advances to the next round, the seed numbers in one game add up to 9 in Rd32, 5 in Sweet16, and 3 in the Elite8. Four regional champions play in the Final4. Teams having the same seed number from different regions are considered as having equal strength. In this situation, a bracket can illustrate better than hundreds of words. (Refer to Figure 3)

1.4. March Madness Betting

March Madness betting refers to wagering on the NCAA Men's Basketball Championships. March Madness gambling is now the second biggest sports betting market in terms of dollars wagered on the heels of the NFL. Besides, the competition has turned into the focus of both fun and serious gambling on the result of these basketball tournaments.

Points are awarded to the correct prediction of each game. Of course, there is no one certain method about how to score his/her pool. Everyone can customize his/her own scoring system, but the overall set up is usually the same.

We apply two types of scoring system: one is the doubling points system, and the other is the simple scoring system. Under the doubling points system: each prediction is awarded one

Figure 1: Diagram of an NCAA regional basketball tournament

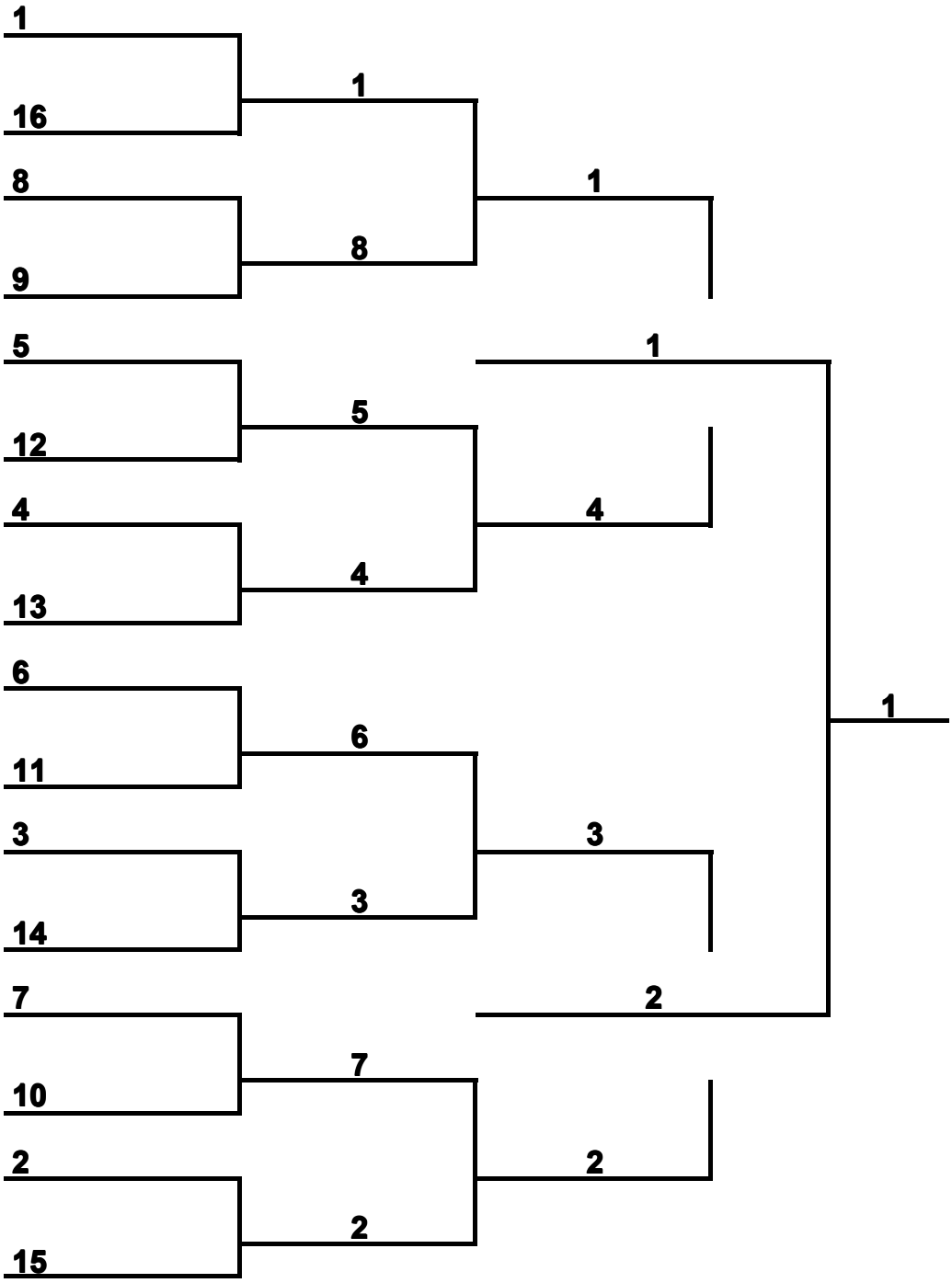


Figure 2: Simple structure of NCAA tournament from 2001 to present

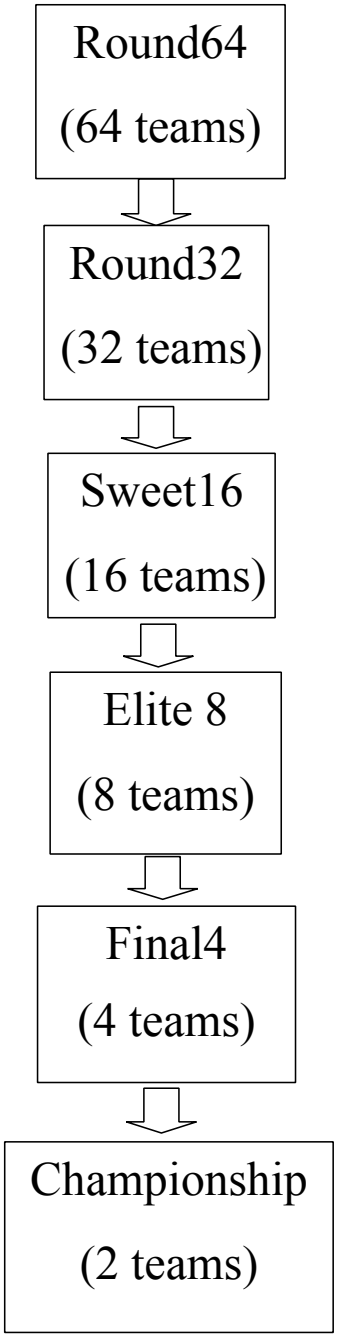
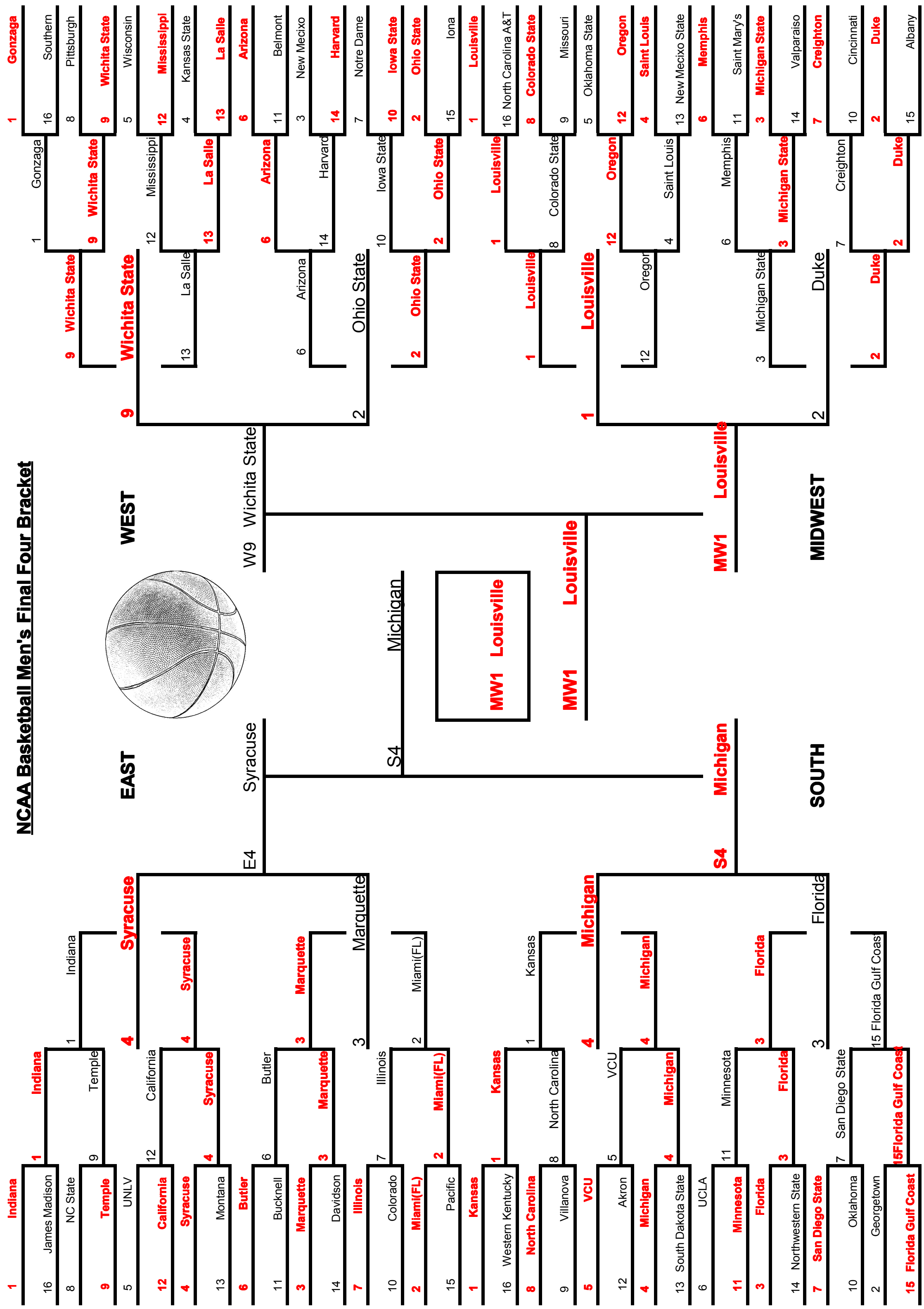


Figure 3: Real NCAA March Madness bracket in the 2012-2013 season (This template is downloaded from: www.samplewords.com/ncaa-blank-printable-tournament-bracket/)



point for correctly picking the winner of a Rd64 game, two points for a Rd32 game, four points for a Sweet16 game, and so on. In this system, one might not care about the individual rounds since predicting the correct champion is worth as much as the entire Rd64 combined. In other words, 1 game in the Championship is equal to 32 games in Rd64. The system puts more weight on later rounds but not the first several rounds. Correct picks in later rounds are awarded more points since picking game winners becomes tougher in later rounds. Under the simple system: each prediction is awarded one point for each game in each round. In this system, one really cares about each game in the whole tournament with no prediction discrimination existing among rounds. Presently, other types of reward bonuses are not taken into account, i.e., upset bonus. Table 1 is the summary of the NCAA Men's Division I Basketball tournament bet scoring system used here.

Table 1: Bet scoring systems

	Rd64	Rd32	Sweet16	Elite8	Final4	Championship	Total
Simple	1	1	1	1	1	1	63
Doubling	1	2	4	8	16	32	192

The point value shown in the table is just for each game, which means the total points is the sum of the number of games per round multiplied by the points per game. No matter which scoring system you use, the winner of the betting is the person who gets the most points.

However, one should first determine which scoring system to be used and plan his/her picks accordingly for winning the bet/pool.

2. BASKETBALL ANALYSIS VARIABLES AND RATING

2.1. Basketball Analysis Variables

At most levels, it's important to record and track stats which help determine the effectiveness of a player or team when they are playing a basketball game. Kubatko et al. (2007) propose some basic basketball statistics. For instance, Field goals attempted, Field goals made, Free throw attempted, Free throw made, Offensive rebounds, Defensive rebounds, Turnovers. These basketball analysis variables observed in a game are referred as the in-game statistics. The in-game statistics reflect the various performances of both competing teams that play in a game and hence provide insight as to how a team shall prepare for the game. However, the in-game statistics can not be used as the predictive variables since they are not observable prior to the game. Unruh (2013) and Long (2013) propose to use the medium of the in-game statistics collected from the recent 3 games to predict the next game outcome. This approach is definitely not applicable in the field of long-term prediction or bracketing before the tournament starts. For instance, one cannot predict the games in Final4 before Rd64, since the recent 3 games for a team in Rd32, Sweet16, and Elite8 has not been played yet when bracketing is pressed. Moreover, the in-game statistics of each team are also affected by the rival in a game, which may not fully reflect the strength of the teams if only 3 recent games are under consideration. Therefore, it's reasonable to consider some basketball analysis variables for a given period, 5 months, as the predictive variables which hopefully truly represent the strength of each team. In this work, three basketball analysis statistics are chosen and applied, which are listed as follow. *Points against* of 5 months subtract from *points for* for the same given period (named as DIFF, abbreviated as D, here after); Assist to turnover ratio of 5 months (named as ATRATIO, abbreviated as ATR, here after); Teams' assists per game of 5 months (named, abbreviated as APG).

2.2. Rating

Other data source for statistical research on basketball are also used, i.e., basketball team rating. For ratings, there are a couple available, used to rank teams in the NCAA basketball

tournament: Ratings Percentage Index (RPI), Pomeroy's ratings (Pomeroy), and Jeff Sagarin computer ratings (Sagrin). Due to knowledge limitations, not all ratings are listed above. There are two rating (RPI and Pomeroy) are used to make brackets later by simply comparing the ratings of two competing teams in each round. Besides, two more analysis variables are selected from two ratings which are Pomeroy and Sagrin, i.e., sagrin proxy for strength of schedule (named as SAGSOS, abbreviated as S, here after); the "Pythagorean winning percentage" (named as Pyth, abbreviated as P, here after).

2.2.1. RPI

RPI is the only rating method used by the NCAA Basketball Selection Committee to pick at-large teams and determine seeds for the NCAA tournament. RPI is only available on ESPN.com, which provides readers with a comprehensive source on all 347 Division I teams.

RPI formula:

$$RPI = 0.25 * WP + 0.50 * OSS + 0.25 * OOSS,$$

where, WP = Division I winning percentage, OSS = Opponent strength of schedule, and OOSS = Opponents' Opponent strength of schedule.

Mike Bobinski, the chair of the Selection Committee, says the RPI won't be the best predictive value in the NCAA basketball tournament in February:
"Interestingly, if we went through that, we were all surprised to see that the RPI actually did end up with the highest level of predictive value and the highest correlation with ultimately success in the tournament. That does not mean we're going to use it more or less this year. It's just a very interesting piece of information."

Therefore, one might pursue other higher-level ranking methods besides RPI.

2.2.2. Pomeroy

The Pomeroy's College Basketball Ratings first published online by Ken Pomeroy in 2003, are a series of predictive ratings of men's college basketball teams. This rating is

derived from the Pythagorean winning percentage (Pyth) which is based on the formula:

$$Pyth = \frac{AdjO^{10.25}}{(AdjO^{10.25} + AdjD^{10.25})},$$

where $AdjO$ is the adjusted offensive efficiency, an estimate of the offensive efficiency (points scored per 100 possessions) that a team would have against the average Division I defense; $AdjD$ is the adjusted defensive efficiency, an estimate of the defensive efficiency (points allowed per 100 possessions) that a team would have against the average Division I offense. For details, please refer to Pomeroy (2012).

Kubatko et al. (2007) figure out the fact that in different fields uses different values for the exponent in the original Pythagorean winning percentage formula:

$$Pyth = \frac{PTS_t^x}{PTS_t^x + PTS_o^x}.$$

where the subscript t indicates team, the subscript o indicates opponent, and the superscript x is an exponent that is empirically determined. Summarily, in the field of baseball, the exponent x of round 1.8 works best. In NBA basketball, the value of x is between about 13 to 17, Oliver has been joined by ESPN in the use of exponent of 16.5. In other leagues, e.g., college, high school and the WNBA, smaller exponent values work better. Especially in NCAA college basketball, PTS_t and PTS_o can be replaced in original formula by $AdjO$ and $AdjD$, and the exponent x is 10.25.

The ratings information of each team in each season is available on kenpom.com.

2.2.3. Sagarin

Jeff Sagarin has been providing ratings for *USA TODAY* since 1985. The ratings method has been used by the Selection Committee to help determine the participating teams in the NCAA basketball tournament since 1984. Exact details for this method are not publicly available, so one cannot know the exact formula. West (2006, 2008) says that this ratings is mainly based on two different ratings: Sagrin's personal modification of the chess rating system and the PURE POINTS method.

3. STATISTICAL MODELS FOR BRACKETING

3.1. Literature Review

For predicting outcome of games in the NCAA tournament, Kvam and Sokol (2006) use a combined logistic regression 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 an ordinal logistic regression model and expectation (Restricted OLRE model) on π_{ik}, the probability of team i has k winnings in the tournament as follows:

$$\pi_{ik} = \frac{\exp(\alpha_k + x_i'\beta)}{1 + \exp(\alpha_k + x_i'\beta)} - \sum_{j=0}^{k-1} \pi_{ij},\ k = 0, \cdots, 6;$$

where α_k is the intercept for k winnings, x_i is a vector of values for team i on the predictor variables, β is a vector of coefficients associated with the predictor variables, and the last term presents the cumulative sum of the probabilities of winning j games ($j = 0, \ldots, k-1$); this term would be equal to 0 for $k = 0$. Putting all π_{ik} ($i = 1, \ldots, 64; k = 0, \ldots, 6$) in a 64×7 matrix, West (2006) requires the sums of each column must be equal to 32, 16, 8, 4, 2, 1, and 1, respectively. West (2006) then applies the method advocated by Lang (2004, 2005) to fit his Restricted OLRE model.

In order to unify all mathematical symbols in this thesis, we rewrite the Restricted OLRE model into another equivalent form, which is a restricted proportional odds model on Z_i, the total number of winning as follows:

$$\begin{cases} P(Z_i \le k) = \dfrac{\exp(\alpha_k + x_i'\beta)}{1 + \exp(\alpha_k + x_i'\beta)},\ k = 0, \ldots, 5; \\ P(Z_i \le 6) = 1. \end{cases} \tag{1}$$

subjected to

$$\sum_{i=1}^{64} P(Z_i \ge k) = 2^{6-k},\ k = 1, \ldots, 6,$$

Note that α_k in the Restricted OLRE model necessarily increase with k, since $P(Z_i \le k)$ is the cumulative distribution function of Z_i.

Actually any probability model must satisfy some basic probability requirements, e.g., probability self-consistency requirement, which is defined as below:

Definition 1. *First let Z_i be the total number of winning games of team i, where Z_i is a random polynomial variable taking integer values in $\{0, 1, \ldots, 6\}$. Let $O_i^{(k)}$ be the index set of all the possible competing teams against team i in the k^{th} round and $U_i^{(k)}$ be the index set of all the possible competing teams against the team i up to the k^{th} round, i.e., $U_i^{(k)} = \bigcup_{j=0}^{k} O_i^{(j)}$ with $O_i^{(0)} = \{i\}$. For instance, $O_2^{(3)} = \{5, 6, 7, 8\}$; $U_2^{(3)} = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Clearly, for $\forall$ $i = 1, 2, \ldots, 64$ and $k = 1, 2, \cdots, 6$, $O_i^{(k)}$ and $U_i^{(k)}$ are fully determined before the start of the Rd64 in the tournament. The so-called probability self-consistency requirement is that $P(Z_i \geq k)$ must satisfy*

$$\sum_{j \in U_i^{(k)}} P\left(Z_j \geq k\right) = 1.$$

This requirement is necessary because of the single-elimination rule adopted in the last 6 rounds of the tournament.

Clearly, the restriction $\sum_{i=1}^{64} P(Z_i \geq k) = 2^{6-k}$ in the Restricted OLRE model is only necessary but not sufficient for the probability self-consistency requirement being observed. Therefore, the Restricted OLRE model doesn't necessarily satisfy the probability self-consistency requirement, which is a fatal flaw of this model. Here we propose a new statistical model for bracketing the NCAA Men's Division I Basketball Tournament.

3.2. Logistic Conditional Probability Model (LCPM)

To bracket the winner of each game in the tournament, it is natural to consider the total number of winning games for each team in a tournament (denoted as Z, here after). Since each game in the tournament has a dichotomous outcome, We may use a binary random variable $I_{ij}^{(k)}$ to denote the result of the game between team i and team j in the k^{th} round of the tournament, i.e.,

$$I_{ij}^{(k)} = \begin{cases} 1, & \text{if } i^{th} \text{ team win;} \\ 0, & \text{if } j^{th} \text{ team win.} \end{cases}$$

We consider a logistic conditional probability model for

$$p_{ij}^{(k)} \triangleq P\left(I_{ij}^{(k)} = 1 | Z_j \geq k-1, Z_i \geq k-1\right).$$

Formally,

$$\log \frac{p_{ij}^{(k)}}{1-p_{ij}^{(k)}} = (x_i - x_j)'\beta^{(k)},\ k = 1, \cdots, 6,$$

where $(x_i - x_j)$ is the vector of the predictor variables spread between team i and team j, and $\beta^{(k)}$ are the associated coefficients in the k^{th} round. These logistic conditional probability models implies

$$\begin{aligned} P(Z_i \geq k) &= P(Z_i \geq k, Z_i \geq k-1) = P(Z_i \geq k-1)P(Z_i \geq k | Z_i \geq k-1) \\ &= P(Z_i \geq k-1) \sum_{j \in O_i^{(k)}} P(I_{ij}^{(k)} = 1, Z_j \geq k-1 | Z_i \geq k-1) \\ &= P(Z_i \geq k-1) \sum_{j \in O_i^{(k)}} P(Z_j \geq k-1) P(I_{ij}^{(k)} = 1 | Z_i \geq k-1, Z_j \geq k-1). \end{aligned}$$

Therefore,

$$P(Z_i \geq k) = P(Z_i \geq k-1) \sum_{j \in O_i^{(k)}} P(Z_j \geq k-1) p_{ij}^{(k)}. \tag{2}$$

Observe $P(Z_i \geq 1) = p_{ij}^{(1)}$, $j \in O_i^{(1)}$, then $P(Z_i \geq k)$ may be computed iteratively via $p_{ij}^{(k)}$, $k = 2, \cdots, 6$. Our model specification implies that it satisfies

$$p_{ij}^{(k)} + p_{ji}^{(k)} = 1,\ \forall\ i, j, k.$$

This makes sense, because in the game played by i^{th} team and j^{th} team, the summation of the probability of i^{th} team wins and the probability of j^{th} team loses in the k^{th} round is 1, and the event that i^{th} team loses is simply the event that j^{th} team wins. By Proposition 1, our logistic conditional probability models automatically satisfy the self-consistency requirement on $P(Z_i \geq k)$. Finally, the expected number of winning games of the i^{th} team in the tournament is

$$EZ_i = \sum_{k=1}^{6} P(Z_i \geq k).$$

Proposition 1. *Every conditional probability model satisfying* $p_{ij}^{(k)} + p_{ji}^{(k)} = 1$ *for* $\forall\ i, j, k$ *must satisfy the probability self-consistency requirement, i.e.,*

$$\sum_{j \in U_i^{(k)}} P(Z_j \geq k) = 1$$

for $\forall\ i, k$.

4. LIVE APPLICATION

4.1. Data Description

In our work, only five basketball covariate variables are utilized, which are DIFF, SAGSOS, ATRATIO, APG, and Pyth. All of these data used in our work are collected from 2002 to 2013.

The data of DIFF is derived from www-personal.umich.edu/ bwest/ (West's homepage). The data SAGSOS is also derived from www-personal.umich.edu/ bwest/ (West's homepage). ATRATIO is an average value for each team in one season and the data is derived from teamrankings.com. APG is also an average value for each team in one season and the data is also derived from teamrankings.com. The data of Pyth is derived from kenpom.com.

4.2. Model Estimating

This analysis first applies the NCAA data set on the new method. Our work uses historical data from the 2002-2003 season through the 2011-2012 season as the training data to test the accuracy for bracketing all 63 games in the 2012-2013 season.

Akaike information criterion with a correction (AICc) is selected as a measure for model selection, since the data size is small. An exhaustive search based on the AICc in the space of all the 2^5 possible candidate models is done to find the best logistic conditional probability model for the games in each round. The preferred model is the one with the minimum AICc value.

The formula for AICc is:

$$AICc = -2\text{logLik} + \frac{2pn}{(n-p)},$$

where -2logLik is the deviance of the statistical model, p is the number of parameters in the statistical model, and n denotes the sample size.

Our logistic conditional probability model is fitted round by round in each season. In the 2012-2013 season, to be specific,

[1] We use the 320 Rd64 games (2002 – 2012) to fit $\beta^{(1)}$ in the conditional $\log \frac{p_{ij}^{(1)}}{1-p_{ij}^{(1)}} = (x_i - x_j)'\beta^{(1)}$ for Rd64. We apply AICc for model selection. Finally, the fitted model for this round is:

$$\log \frac{p_{ij}^{(1)}}{1 - p_{ij}^{(1)}} = -0.004\left(D_i - D_j\right) - 0.164\left(S_i - S_j\right) + 16.727\left(P_i - P_j\right).$$

[2] We use the 160 Rd32 games (2002 – 2012) to fit $\beta^{(2)}$ in the conditional $\log \frac{p_{ij}^{(2)}}{1-p_{ij}^{(2)}} = (x_i - x_j)'\beta^{(2)}$ for Rd32. We apply AICc for model selection. Finally, the fitted model for this round is:

$$\log \frac{p_{ij}^{(2)}}{1 - p_{ij}^{(2)}} = 18.387\left(P_i - P_j\right).$$

[3] We use the 80 Sweet16 games (2002 – 2012) to fit $\beta^{(3)}$ in the conditional $\log \frac{p_{ij}^{(3)}}{1-p_{ij}^{(3)}} = (x_i - x_j)'\beta^{(3)}$ for Sweet16. We apply AICc for model selection. Finally, the fitted model for this round is:

$$\log \frac{p_{ij}^{(3)}}{1 - p_{ij}^{(3)}} = 25.633\left(P_i - P_j\right).$$

[4] We use the 40 Elite8 games (2002 – 2012) to fit $\beta^{(4)}$ in the conditional $\log \frac{p_{ij}^{(4)}}{1-p_{ij}^{(4)}} = (x_i - x_j)'\beta^{(4)}$ for Elite8. We apply AICc for model selection. Finally, the fitted model for this round is:

$$\log \frac{p_{ij}^{(4)}}{1 - p_{ij}^{(4)}} = 12.3\left(P_i - P_j\right).$$

[5] We use the 20 Final4 games (2002 – 2012) to fit $\beta^{(5)}$ in the conditional $\log \frac{p_{ij}^{(5)}}{1-p_{ij}^{(5)}} = (x_i - x_j)'\beta^{(5)}$ for Final4. We apply AICc for model selection. Finally, the fitted model for this round is:

$$\log \frac{p_{ij}^{(5)}}{1 - p_{ij}^{(5)}} = 0.629\left(APG_i - APG_j\right) + 69.335\left(P_i - P_j\right).$$

[6] We use the 10 Championship games (2002 – 2012) to fit $\beta^{(6)}$ in the conditional $\log \frac{p_{ij}^{(6)}}{1-p_{ij}^{(6)}} = (x_i - x_j)'\beta^{(6)}$ for Champion. We apply AICc for model selection. Finally, the fitted model for this round is:

$$\log \frac{p_{ij}^{(6)}}{1 - p_{ij}^{(6)}} = 50.88\left(P_i - P_j\right).$$

Therefore, one may use all these fitted models to compute $p_{ij}^{(k)}$ for $\forall i, j, k$.

Table 2 and Table 3 are the summary of Estimated coefficients and standard errors for two types of statistical models in the 2012-2013 season, respectively.

Table 2: Estimates for LCPM selected by AICc, standard error given in parenthesis (in the 2012-2013 season)

	DIFF	SAGSOS	ATRATIO	APG	Pyth
Rd64	−.004 (.002)	−.164 (.071)			16.727 (3.523)
Rd32					18.387 (3.029)
Sweet16					25.633 (6.298)
Elite8					12.300 (8.520)
Final4				.629 (.390)	69.335 (36.027)
Championship					50.880 (36.370)

Table 3: Estimates for Restricted OLRE model, standard errors given in parenthesis (in the 2012-2013 season)

DIFF	SAGSOS	ATRATIO	APG	Pyth	
−.006 (.001)	−.156 (.050)	−.564 (.600)	.129 (.064)	−10.144 (2.487)	
α_0	α_1	α_2	α_3	α_4	α_5
20.761 (2.771)	22.437 (2.795)	23.609 (2.811)	24.557 (2.811)	25.392 (2.834)	26.173 (2.847)

4.3. Bracketing

Participants in the March Madness Bracket pool need to bet on the results of all 63 games prior to Rd64 of the tournament. This prediction is often referred to as a bracket.

This pool typically assigns a point value to each game in the tournament. The winner of the pool is the person who amasses the maximum points. One may have a couple of strategies to pick the winner for each individual game. In other words, with his filled out bracket, the more points one gets the better. To be honest, there is no way one can make a perfect bracket prediction for one season tournament by using any strategy. However, there are some popular ratings of college basketball teams who make the predictions year by year.

In this thesis, four bracketing methods are mainly mentioned, which are LCPM, Restricted OLRE model, Pomeroy, and RPI.

While doing the LCPM brackets, first of all, all $P(Z_i \geq k)$ for $\forall\ i, k$ can be computed iteratively via $p_{ij}^{(k)}$, i.e.,

$$\begin{cases} P(Z_i \geq 1) = p_{ij}^{(1)}, j \in O_i^{(1)}; \\ P(Z_i \geq k) = P(Z_i \geq k-1) \sum\limits_{j \in O_i^{(k)}} P(Z_j \geq k-1) p_{ij}^{(k)}, k = 2, \cdots, 6. \end{cases} \tag{3}$$

By $Eq(3)$, all of the $P(Z_i \geq k)$ may be computed one by one. Please refer to Figure 4.

In Rd64, in order to pick the winner, one just compares $P(Z_i \geq 1)$ and $P(Z_j \geq 1)$ for teams in $U_i^{(1)}$. For example, in $U_3^{(1)} = \{3, 4\}$, one may pick the winner with the higher probability between $P(Z_3 \geq 1)$ and $P(Z_4 \geq 1)$. Since $P(Z_3 \geq 1) = 0.4758$, $P(Z_4 \geq 1) = 0.5242$, thus the winner in Rd64 is team 3 (Colorado State).

In Rd32, similarly, for each i, one and only one team in $U_i^{(2)}$ may advance to Sweet16. So, We simply compare $P(Z_j \geq 2)$ for $\forall\, j \in U_i^{(2)}$. For example, in $U_3^{(2)} = \{1, 2, 3, 4\}$, one may pick the winner with the highest probability among $P(Z_1 \geq 2)$, $P(Z_2 \geq 2)$, $P(Z_3 \geq 2)$, $P(Z_4 \geq 2)$. In this situation, $P(Z_1 \geq 2) = 0.8628$, $P(Z_2 \geq 2) = 0.0000$, $P(Z_3 \geq 2) = 0.0550$, $P(Z_4 \geq 2) = 0.0822$. The winner of them in Rd32 is team 1 (Louisville) with highest probability 0.8628.

In Sweet16, similarly, for each i, one and only one team in $U_i^{(3)}$ may advance to Elite8. So, We simply compare $P(Z_j \geq 3)$ for $\forall\, j \in U_i^{(3)}$. For example, in $U_3^{(3)} = \{1, 2, 3, 4, 5, 6, 7, 8\}$, one may pick the winner with the highest probability among $P(Z_1 \geq 3)$, $P(Z_2 \geq 3)$, $P(Z_3 \geq 3)$, $P(Z_4 \geq 3)$, $P(Z_5 \geq 3)$, $P(Z_6 \geq 3)$, $P(Z_7 \geq 3)$, $P(Z_8 \geq 3)$. In this condition, $P(Z_1 \geq 3) = 0.7748$, $P(Z_2 \geq 3) = 0.0000$, $P(Z_3 \geq 3) = 0.0191$, $P(Z_4 \geq 3) = 0.0382$, $P(Z_5 \geq 3) = 0.0279$, $P(Z_6 \geq 3) = 0.0283$, $P(Z_7 \geq 3) = 0.1116$, $P(Z_8 \geq 3) = 0.0000$. So, the winner is team 1 (Louisville) with highest probability 0.7748 in Sweet16.

In Elite8, similarly, for each i, one and only one team in $U_i^{(4)}$ may advance to Final4. So, We simply compare $P(Z_j \geq 4)$ for $\forall\, j \in U_i^{(4)}$.

Figure 4: LCPM probability matrix in the 2012-2013 season

TEAM	P1	P2	P3	P4	P5	P6
1 Louisville	0.9985	0.8628	0.7748	0.4882	0.3750	0.3000
2 N Carolina A&T	0.0015	0.0000	0.0000	0.0000	0.0000	0.0000
3 Colorado St	0.4758	0.0550	0.0191	0.0059	0.0000	0.0000
4 Missouri	0.5242	0.0822	0.0382	0.0137	0.0004	0.0000
5 Oklahoma St.	0.4266	0.1932	0.0279	0.0099	0.0001	0.0000
6 Oregon	0.5734	0.2354	0.0283	0.0092	0.0001	0.0000
7 Saint Louis	0.9378	0.5699	0.1116	0.0451	0.0018	0.0003
8 New Mexico St	0.0622	0.0015	0.0000	0.0000	0.0000	0.0000
9 Memphis	0.3284	0.0887	0.0099	0.0020	0.0000	0.0000
10 Saint Mary's(CA)	0.6716	0.2747	0.0757	0.0236	0.0010	0.0001
11 Michigan State	0.8137	0.6143	0.3290	0.1454	0.0285	0.0118
12 Valparaiso	0.1863	0.0223	0.0004	0.0000	0.0000	0.0000
13 Creighton	0.7077	0.2285	0.0961	0.0326	0.0060	0.0008
14 Cincinnati	0.2923	0.0443	0.0077	0.0016	0.0000	0.0000
15 Duke	0.9752	0.7271	0.4811	0.2228	0.0814	0.0400
16 Albany	0.0248	0.0001	0.0000	0.0000	0.0000	0.0000
17 Kansas	0.9922	0.8246	0.4306	0.1993	0.0907	0.0395
18 Western Kentucky	0.0078	0.0000	0.0000	0.0000	0.0000	0.0000
19 North Carolina	0.5559	0.1216	0.0209	0.0057	0.0005	0.0000
20 Villanova	0.4441	0.0537	0.0042	0.0008	0.0000	0.0000
21 VCU	0.8354	0.2857	0.1060	0.0383	0.0056	0.0009
22 Akron	0.1646	0.0115	0.0002	0.0000	0.0000	0.0000
23 Michigan	0.9423	0.7009	0.4381	0.2158	0.0985	0.0534
24 South Dakota State	0.0577	0.0019	0.0000	0.0000	0.0000	0.0000
25 UCLA	0.3177	0.0238	0.0040	0.0008	0.0000	0.0000
26 Minnesota	0.6823	0.1410	0.0675	0.0247	0.0028	0.0003
27 Florida	0.9834	0.8351	0.7076	0.4226	0.2888	0.2179
28 Northwestern St.	0.0166	0.0001	0.0000	0.0000	0.0000	0.0000
29 San Diego State	0.6635	0.2120	0.0194	0.0050	0.0000	0.0000
30 Oklahoma	0.3365	0.0689	0.0023	0.0004	0.0000	0.0000
31 Georgetown	0.8829	0.7122	0.1992	0.0865	0.0185	0.0043
32 Florida Gulf Coast	0.1171	0.0069	0.0000	0.0000	0.0000	0.0000
33 Indiana	0.9878	0.9082	0.5928	0.4086	0.3354	0.1608
34 James Madison	0.0122	0.0000	0.0000	0.0000	0.0000	0.0000
35 NC State	0.5923	0.0725	0.0102	0.0037	0.0002	0.0000
36 Temple	0.4077	0.0193	0.0008	0.0002	0.0000	0.0000
37 UNLV	0.5276	0.0647	0.0049	0.0016	0.0001	0.0000
38 California	0.4724	0.0346	0.0014	0.0003	0.0000	0.0000
39 Syracuse	0.9914	0.9007	0.3899	0.2493	0.1585	0.0463
40 Montana	0.0086	0.0000	0.0000	0.0000	0.0000	0.0000
41 Butler	0.5742	0.2435	0.0366	0.0063	0.0000	0.0000
42 Bucknell	0.4258	0.0972	0.0044	0.0004	0.0000	0.0000
43 Marquette	0.6189	0.4857	0.1685	0.0464	0.0057	0.0001
44 Davidson	0.3811	0.1735	0.0130	0.0016	0.0000	0.0000
45 Illinois	0.5725	0.1345	0.0748	0.0164	0.0001	0.0000
46 Colorado	0.4275	0.0696	0.0294	0.0050	0.0000	0.0000
47 Miami(FL)	0.9889	0.7959	0.6732	0.2603	0.0607	0.0066
48 Pacific	0.0111	0.0000	0.0000	0.0000	0.0000	0.0000
49 Gonzaga	0.9877	0.6443	0.4975	0.3443	0.2365	0.0871
50 Southern	0.0123	0.0000	0.0000	0.0000	0.0000	0.0000
51 Pittsburgh	0.4331	0.1874	0.1317	0.0854	0.0563	0.0127
52 Wichita State	0.5669	0.1683	0.0913	0.0505	0.0129	0.0008
53 Wisconsin	0.5827	0.4848	0.2177	0.1393	0.0561	0.0114
54 Mississippi	0.4173	0.2494	0.0335	0.0147	0.0008	0.0000
55 Kansas State	0.5609	0.1888	0.0251	0.0110	0.0017	0.0000
56 La Salle	0.4391	0.0771	0.0032	0.0009	0.0000	0.0000
57 Arizona	0.7141	0.4866	0.3620	0.1537	0.0459	0.0041
58 Belmont	0.2859	0.0881	0.0185	0.0035	0.0001	0.0000
59 New Mexico	0.9162	0.4238	0.2323	0.0762	0.0114	0.0002
60 Harvard	0.0838	0.0015	0.0000	0.0000	0.0000	0.0000
61 Notre Dame	0.4331	0.2139	0.0662	0.0180	0.0026	0.0000
62 Iowa State	0.5669	0.3229	0.1324	0.0423	0.0089	0.0002
63 Ohio State	0.8556	0.4601	0.1887	0.0602	0.0057	0.0001
64 Iona	0.1444	0.0031	0.0000	0.0000	0.0000	0.0000

In Final4, similarly, for each i, one and only one team in $U_i^{(5)}$ may advance to Champion. So, We simply compare $P(Z_j \geq 5)$ for $\forall j \in U_i^{(5)}$.

In Championship, similarly, for each i, one and only one team in $U_i^{(6)}$ can be the championship team. So, We simply compare $P(Z_j \geq 6)$ for $\forall j \in U_i^{(6)}$.

For Restricted OLRE model, pick the winners by using the same way as LCPM. Please refer to Figure 5 to see all the probabilities by Restricted OLRE model.

For Pomeroy, the winner in each game is determined by simply comparing the Pomeroy's ratings of the two competing teams in $O_i^{(k)}$ in each round.

For RPI, the winner in each game is also determined by simply comparing the RPI ratings of the two competing teams in $O_i^{(k)}$ in each round.

The work applies these four bracketing methods into practice in the 2012-2013 season, which is refer to as one case. Then four brackets are shown in Figure 6 through Figure 9. The teams with red color in four brackets are the winners, which are successfully predicted in the 2012-2013 season.

4.4. March Madness Scoring

Table 4 shows how the four strategies would have performed in the 2013 March Madness bracketing, respectively. Here, the prediction accuracy is the ratio of the number of games correctly predicted in one season divided by the total number of games in a tournament (63 games). The value in the 3^{rd} line is the ratio of the total points gained in a bracket divided by the total possible points (63 points) in a bracket, under the simple scoring system. The value in the 4^{th} line is the ratio of the total points gained in a bracket divided by the total possible points (192 points) in a bracket, under the doubling points system.

Table 4: The summary of scoring performance and prediction accuracy for four brackets in the 2012-2013 season

	LCPM	Restricted OLRE model	Pomeroy	RPI
Correctly Predicted	43	40	42	40
Prediction Accuracy	68.3%	63.5%	66.7%	63.5%
Simple System	68.3%	63.5%	66.7%	63.5%
Doubling System	62.5%	38.5%	63.5%	32.8%

Figure 5: Restricted OLRE model probability matrix in the 2012-2013 season

TEAM	Q1	Q2	Q3	Q4	Q5	Q6
1 Louisville	0.9646	0.8438	0.6379	0.4106	0.2349	0.1257
2 N Carolina A&T	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000
3 Colorado St	0.6615	0.2629	0.0974	0.0391	0.0171	0.0080
4 Missouri	0.7243	0.3263	0.1284	0.0527	0.0233	0.0109
5 Oklahoma St.	0.7427	0.3480	0.1401	0.0581	0.0258	0.0121
6 Oregon	0.5715	0.1939	0.0674	0.0264	0.0114	0.0053
7 Saint Louis	0.7291	0.3319	0.1314	0.0541	0.0240	0.0112
8 New Mexico St	0.0716	0.0132	0.0040	0.0016	0.0007	0.0003
9 Memphis	0.4661	0.1347	0.0444	0.0171	0.0074	0.0034
10 Saint Mary's(CA)	0.6940	0.2938	0.1120	0.0455	0.0200	0.0094
11 Michigan State	0.8466	0.5108	0.2449	0.1102	0.0507	0.0241
12 Valparaiso	0.1876	0.0385	0.0117	0.0044	0.0019	0.0008
13 Creighton	0.7551	0.3636	0.1486	0.0621	0.0276	0.0129
14 Cincinnati	0.6151	0.2248	0.0804	0.0319	0.0139	0.0065
15 Duke	0.9204	0.6913	0.4156	0.2160	0.1070	0.0527
16 Albany	0.0080	0.0013	0.0003	0.0001	0.0000	0.0000
17 Kansas	0.9053	0.6484	0.3679	0.1834	0.0888	0.0432
18 Western Kentucky	0.0025	0.0004	0.0001	0.0000	0.0000	0.0000
19 North Carolina	0.5705	0.1932	0.0671	0.0263	0.0114	0.0053
20 Villanova	0.3548	0.0883	0.0280	0.0107	0.0045	0.0021
21 VCU	0.8076	0.4402	0.1950	0.0843	0.0382	0.0180
22 Akron	0.1829	0.0374	0.0114	0.0042	0.0018	0.0008
23 Michigan	0.9129	0.6698	0.3912	0.1990	0.0974	0.0477
24 South Dakota State	0.0440	0.0078	0.0023	0.0008	0.0004	0.0002
25 UCLA	0.4022	0.1065	0.0343	0.0131	0.0056	0.0026
26 Minnesota	0.6780	0.2782	0.1045	0.0421	0.0185	0.0086
27 Florida	0.9664	0.8513	0.6515	0.4253	0.2462	0.1327
28 Northwestern St.	0.0087	0.0015	0.0004	0.0002	0.0001	0.0000
29 San Diego State	0.5708	0.1934	0.0672	0.0263	0.0114	0.0053
30 Oklahoma	0.3818	0.0984	0.0315	0.0120	0.0052	0.0024
31 Georgetown	0.7585	0.3681	0.1512	0.0632	0.0281	0.0132
32 Florida Gulf Coast	0.0360	0.0064	0.0019	0.0007	0.0003	0.0002
33 Indiana	0.9654	0.8476	0.6447	0.4179	0.2405	0.1292
34 James Madison	0.0032	0.0005	0.0001	0.0000	0.0000	0.0000
35 NC State	0.5642	0.1891	0.0654	0.0256	0.0111	0.0051
36 Temple	0.2247	0.0481	0.0148	0.0056	0.0024	0.0011
37 UNLV	0.4302	0.1183	0.0384	0.0148	0.0064	0.0030
38 California	0.2336	0.0504	0.0155	0.0059	0.0025	0.0011
39 Syracuse	0.9004	0.6350	0.3542	0.1745	0.0840	0.0408
40 Montana	0.0046	0.0008	0.0002	0.0001	0.0000	0.0000
41 Butler	0.4036	0.1070	0.0345	0.0132	0.0057	0.0026
42 Bucknell	0.2563	0.0567	0.0175	0.0066	0.0028	0.0013
43 Marquette	0.5825	0.2013	0.0704	0.0277	0.0120	0.0056
44 Davidson	0.2962	0.0688	0.0215	0.0082	0.0035	0.0016
45 Illinois	0.5279	0.1671	0.0568	0.0221	0.0096	0.0044
46 Colorado	0.4349	0.1203	0.0392	0.0151	0.0065	0.0030
47 Miami(FL)	0.8337	0.4859	0.2264	0.1004	0.0459	0.0217
48 Pacific	0.0082	0.0013	0.0003	0.0001	0.0000	0.0000
49 Gonzaga	0.9343	0.7354	0.4706	0.2570	0.1311	0.0656
50 Southern	0.0060	0.0010	0.0002	0.0001	0.0000	0.0000
51 Pittsburgh	0.8733	0.5679	0.2912	0.1361	0.0637	0.0306
52 Wichita State	0.6538	0.2561	0.0943	0.0377	0.0165	0.0077
53 Wisconsin	0.8982	0.6288	0.3479	0.1705	0.0818	0.0397
54 Mississippi	0.6876	0.2874	0.1089	0.0441	0.0194	0.0090
55 Kansas State	0.5790	0.1989	0.0694	0.0273	0.0119	0.0055
56 La Salle	0.3265	0.0786	0.0247	0.0094	0.0040	0.0018
57 Arizona	0.7643	0.3759	0.1556	0.0653	0.0292	0.0137
58 Belmont	0.4121	0.1105	0.0357	0.0137	0.0059	0.0027
59 New Mexico	0.6267	0.2337	0.0843	0.0335	0.0146	0.0068
60 Harvard	0.0343	0.0061	0.0018	0.0007	0.0003	0.0002
61 Notre Dame	0.5323	0.1697	0.0577	0.0225	0.0097	0.0045
62 Iowa State	0.5957	0.2104	0.0742	0.0293	0.0127	0.0059
63 Ohio State	0.8221	0.4649	0.2116	0.0927	0.0421	0.0199
64 Iona	0.0421	0.0075	0.0022	0.0008	0.0004	0.0002

Figure 6: LCPM March Madness bracket in the 2012-2013 season

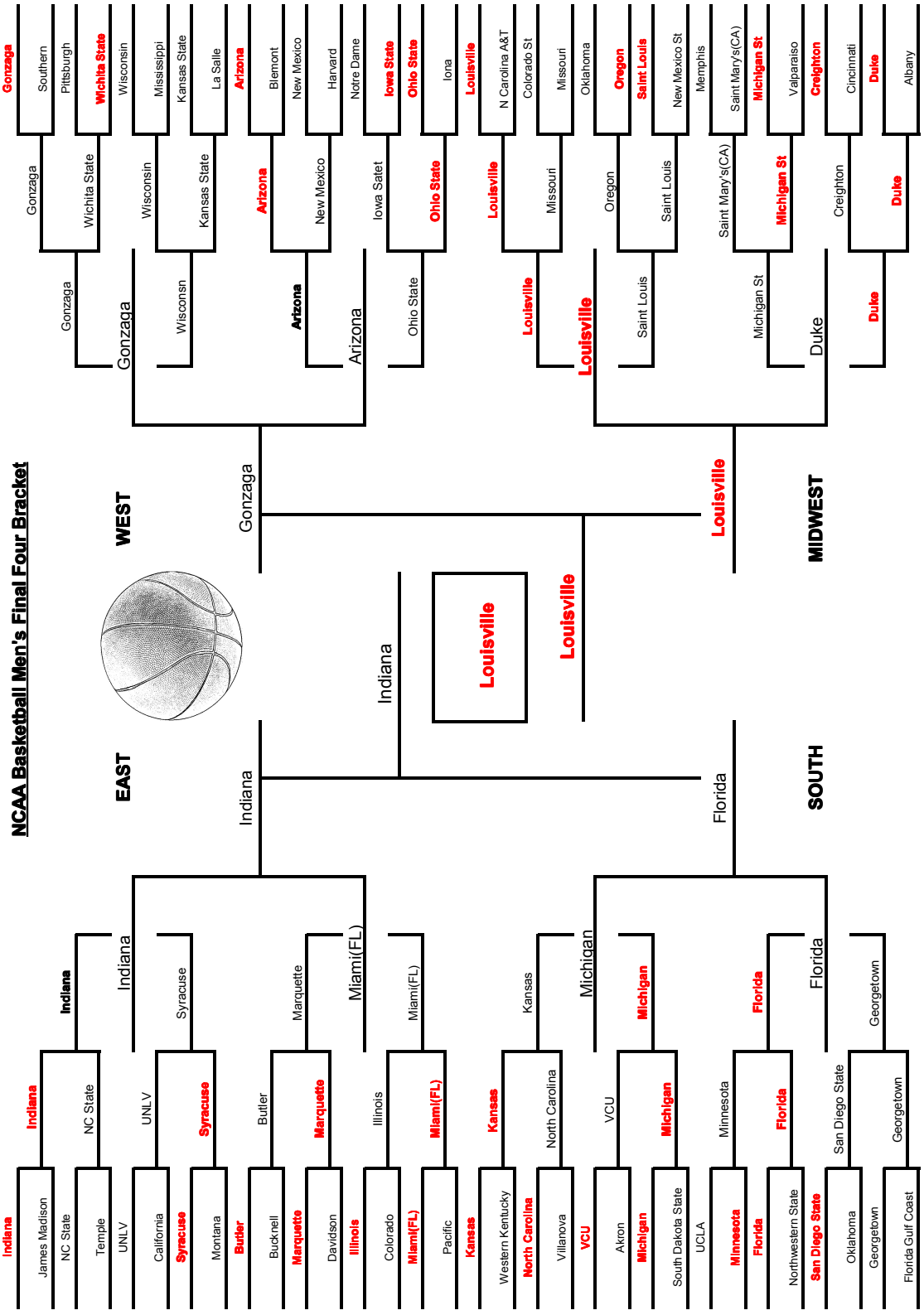


Figure 7: Restricted OLRE model March Madness bracket in the 2012-2013 season

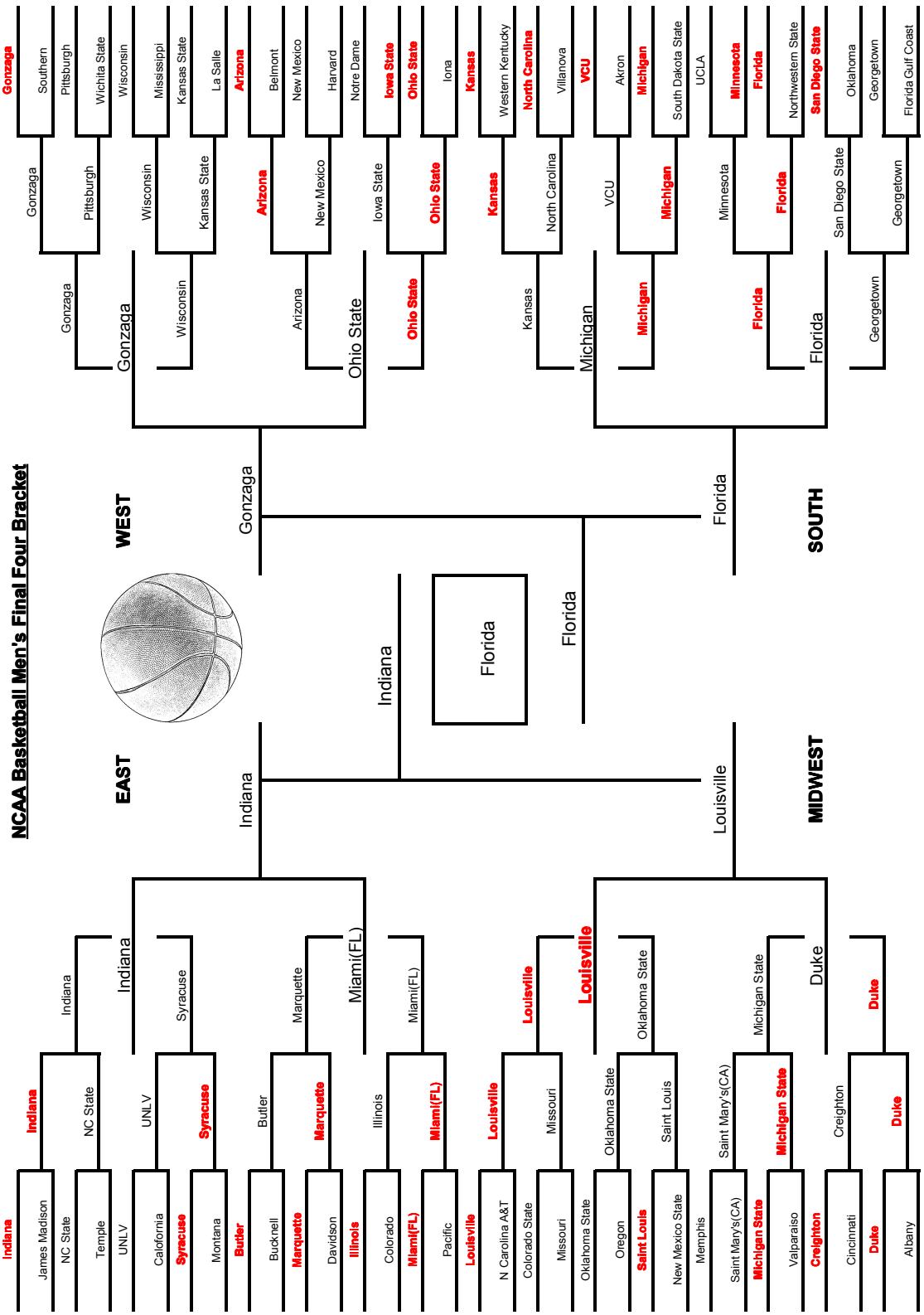


Figure 8: Pomeroy March Madness bracket in the 2012-2013 season

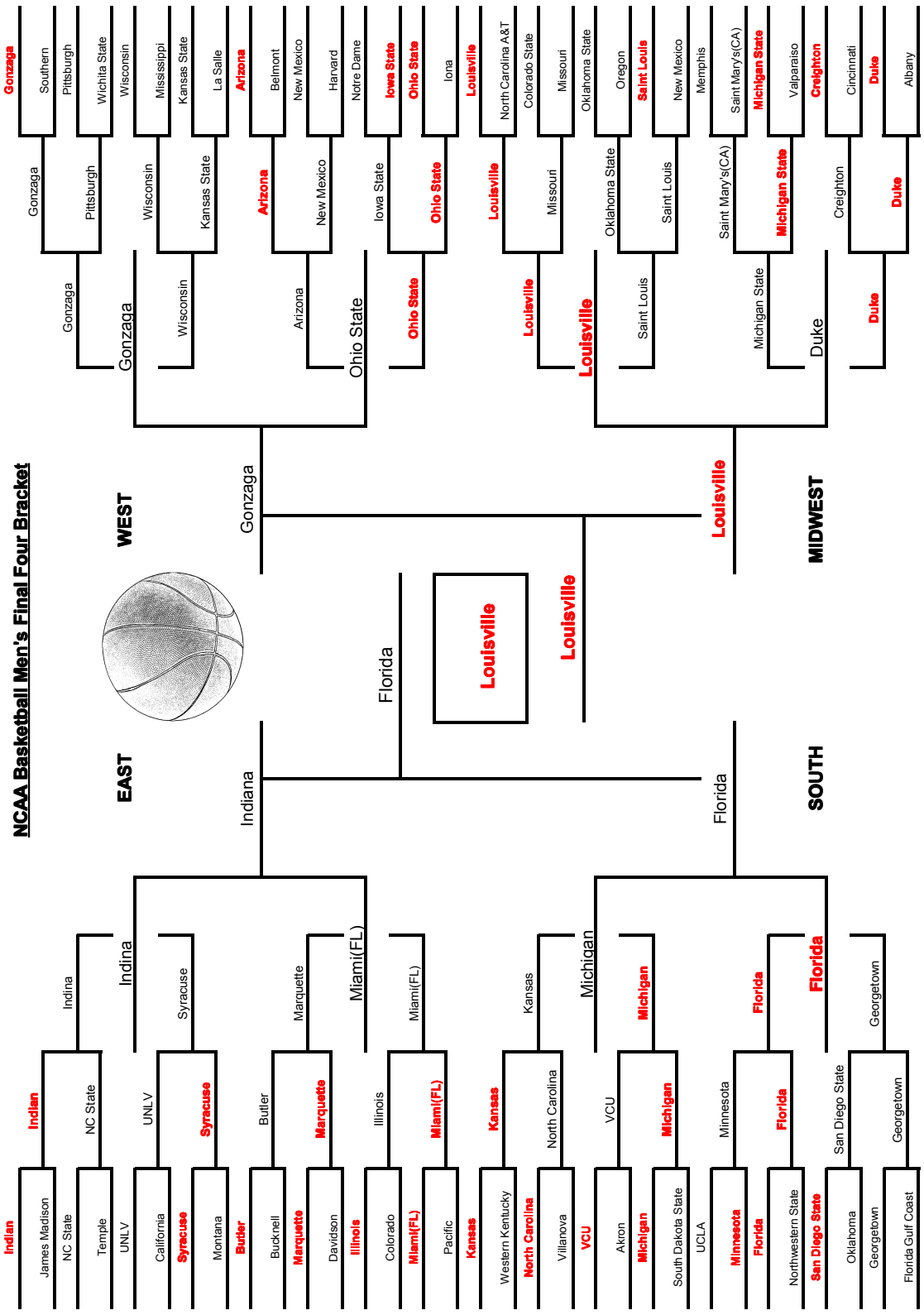
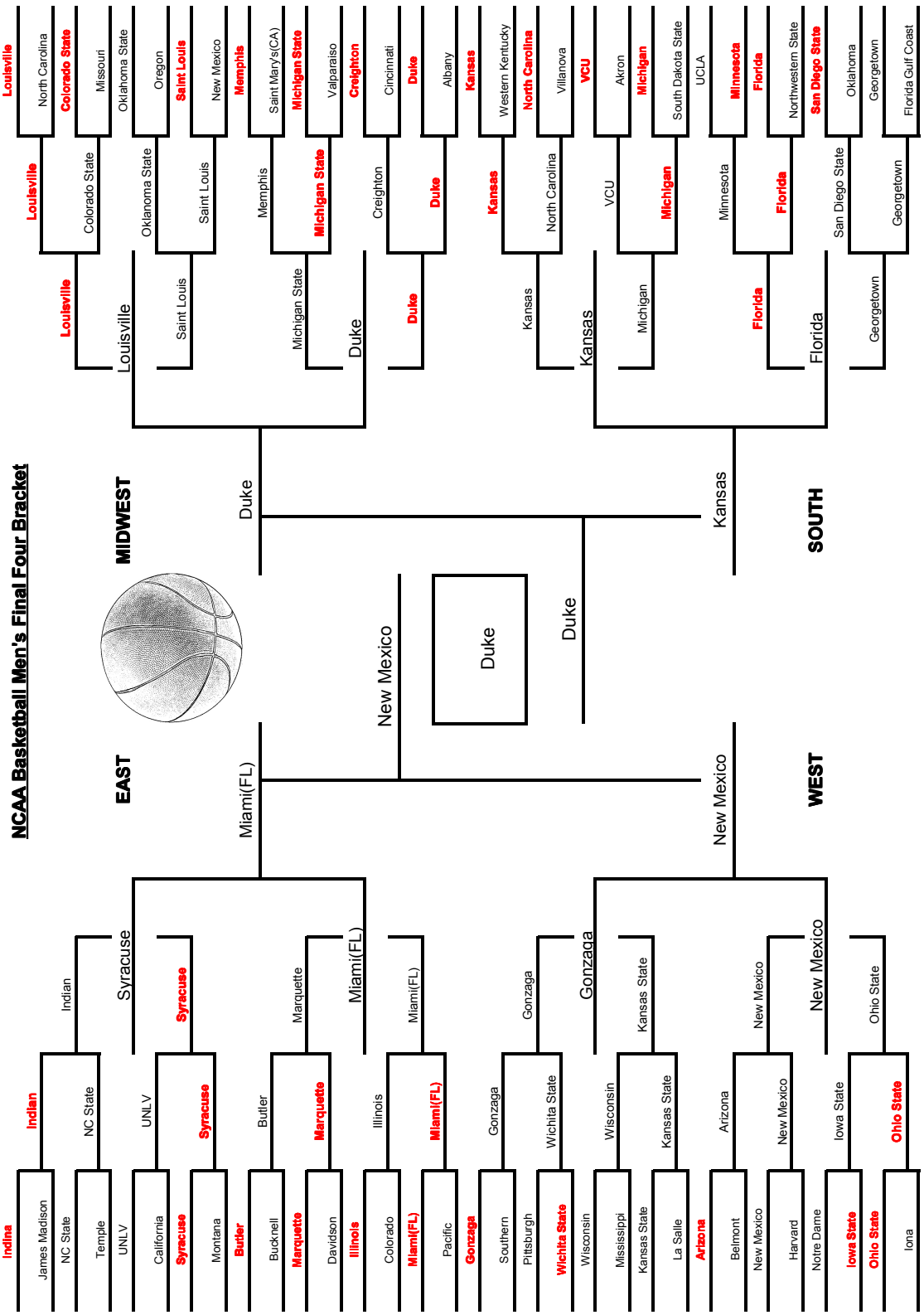


Figure 9: RPI March Madness bracket in the 2012-2013 season



In the 2012-2013 season, the LCPM approach indeed performs better than the Restricted OLRE model, regarding either the prediction accuracy or scoring performance. If the doubling points system is used as the measurement, LCPM totally beats Restricted OLRE model (62.5% versus 38.5%). LCPM also successfully predicts 43 games, while Restricted OLRE model only successfully predicts 40 games. In the first four rounds (Rd64, Rd32, Sweet16, Elite8), the two approaches perform similarly; however, in the last two rounds (Final4 and Championship), Restricted OLRE model doesn't predict one game successfully (LCPM successfully predicted two games). Everyone knows that it gets more difficult to pick winners in later rounds as more and more teams are eliminated; therefore, LCPM's performance in the last two rounds is definitely outstanding. Please refer to Table 5 for details regarding numbers of picking winners correctly. Hence, if LCPM and Restricted OLRE model are given to fill in one bracket, LCPM approach is apparently preferred, since it can make a better NCAA Men's College Basketball tournament bracket. This can make it more likely to win the March Madness, and even lead to the huge awards of millions of dollars offered by others (i.e., Papa John's, American Online, Sportsbook.com, etc.).

Table 5: Prediction performance in the 2012-2013 season

	Rd64	Rd32	Sweet16	Elite8	Final4	Championship
LCPM	24	12	4	1	1	1
Restricted OLRE model	22	12	5	1	0	0

4.5. The Summary of Prediction Accuracy in Recent Three Seasons

Actually, by using the same idea, We also do two more cases in the 2011-2012 season and in the 2010-2011 season, including model estimating, bracketing, March Madness scoring, computing prediction accuracy, etc. But the details for all of them are not displayed one by one in order to avoid redundancy in this thesis. In the 2011-2012 season, our work uses historical data from the 2002-2003 season to the 2010-2011 season as the training data to test the accuracy for bracketing all 63 games in the 2011-2012 season. Similarly, in the 2010-2011 season, our work uses historical data from the 2002-2003 season to the 2009-2010 season as the training data to test the accuracy for bracketing all 63 games in the 2010-2011 season.

Table 6: The summary of the prediction accuracy from the 2010-2011 season through the 2012-2013 season

Season	LCPM	Restricted OLRE model	Pomeroy	RPI
2010 – 2011	60.3%	58.7%	58.7%	49.2%
2011 – 2012	74.6%	69.8%	69.8%	57.1%
2012 – 2013	68.3%	63.5%	66.7%	63.5%
Average	67.7%	64.0%	65.1%	56.6%

Table 6 only shows the summary of all prediction accuracy generated by four methods season by season, from the 2010-2011 season through the 2012-2013 season. Compared with all accuracy performances, LCPM indeed results in a superior performance to the Restricted OLRE model at bracketing the NCAA Men's Division I basketball tournament in the recent three seasons. Among all four methods, only LCPM always has the ideal prediction result in each season. Especially in the 2011-2012 season, LCPM almost got 75% prediction accuracy. The average prediction accuracy is displayed as follows in descending order: 67.7%, 65.1%, 64.0%, 56.6%. Currently, we only use five covariate variables for model estimating, which might not be the optimal combination of predictor variables. However, if one can use a perfect combination of predictor variables, LCPM's approach may perform much better.

5. DISCUSSION

In terms of bracketing the tournament, LCPM outperforms the other methods (Restricted OLRE model, Pomeroy, RPI) in the recent three seasons. Summarily, the number of games successfully predicted by LCPM are always larger than those by three other methods in the recent three seasons. In the recent three seasons, LCPM successfully predicts 128 games out of 189 games, the Restricted OLRE model has 121 games, Pomeroy has 123 games, and RPI has 107 games. Technically LCPM has more advantages than Restricted OLRE model. The Restricted OLRE model is exclusively subjected to the basic requirement $\sum_{i=1}^{64} P(Z_i \geq k) = 2^{6-k}$ (West (2006, 2008) is adjusted to satisfy the constraint), but not sufficient to satisfy the probability self-consistency requirement. In contrast, LCPM satisfies both of the requirements. In addition, my bracketing does not need any simulation to get the final result. However, the Restricted OLRE model needs the simulation. Simulation is computationally expensive and takes time to run. Simulation number is often at the level of 10,000 times, or even bigger. Thus, the Restricted OLRE model performs ineffectually compared to LCPM. Both LCPM and Restricted OLRE model methods are flexible, i.e., the two statistical models are flexible to incorporate additional team-level predictors of success in the tournament. On the other hand, RPI and Pomeroy, of which the formulas are inflexible, cannot incorporate other information. In addition, our bracketing approach also has one more merit. One may apply other link functions by using our approach besides logit link function $g(x) = log\frac{x}{1-x}$, e.g., probit link, as long as the link function $g(x)$ satisfies the requirement $g^{-1}(x) + g^{-1}(-x) = 1$. In the future research, one might attempt to combine other success predictors or try other link functions to better fit this model to possibly improve its predictive power.

REFERENCES

[1] Breiter, D. and Carlin, B. (1997). *How to Play Office Pools If You Must*. CHANCE, **10**, 1, p5-11.

[2] Bialik, C. (2006). *Picking the Perfect NCAA Bracket*. The Wall Street Journal, p1-2.

[3] Brown, M. and Sokol, J. (2010). *An Improved LRMC Method for NCAA Basketball Prediction*. Journal of Quantitative Analysis in Sports, **6**, 3, p1-3.

[4] Kvam, P.S. and Sokol, J. (2006). *A Logistic Regression/Markov Chain Model for NCAA Basketball*. Accepted for Publication in Naval Research Logistics 53, p1-3.

[5] Kubatko, J. and Oliver, D. and Pelton, K. and Rosenbaum, D.T. (2007). *A Starting Point for Analyzing Basketball Statistics*. Journal of Quantitative Analysis in Sports, **3**, 3, p1-18.

[6] Lang, J.B. (2004). *Multinomial-Poisson Homogeneous (MPH) and Homogeneous Linear Predictor (HLP) Models: A Brief Description*. Online HTML Document www.stat.uiowa.edu/ jblang/mph.fitting/mph.hlp.description.htm

[7] Lang, J.B. (2005). *Homogeneous Linear Predictor Models for Contingency Tables*. JASA (T & M), 100, p121-134.

[8] Long, J. (2013). *Development of a Prediction Model for the NCAA Division-I Football Championship Subdivision*. Unpublished Master's Thesis Paper, Department of Statistics, North Dakota State University.

[9] Unruh, S. (2013). *Analysis of Significant Factors in Division I Men's College Basketball and Development of a Predictive Model*. Unpublished Master's Thesis Paper, Department of Statistics, North Dakota State University.

[10] 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, p3-8.

[11] West, B.T. (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, p6-8.

APPENDIX

A.1. Technical Details

Proof. This proposition is proved by induction.

1. When $k = 0$, $U_i^{(0)} = i$. Then $\sum_{j \in U_i^{(k)}} P(Z_j \geq k) = P(Z_i \geq 0) = 1$. The result is trivially true.

2. Suppose the result holds when $k = s$, i.e., $\sum_{j \in U_i^{(s)}} P(Z_j \geq s) = 1$. Then by Eq(2).

$$\sum_{j \in U_i^{(s+1)}} P(Z_j \geq s+1) = \sum_{j \in U_i^{(s+1)}} P(Z_j \geq s) \sum_{l \in O_j^{(s+1)}} P(Z_l \geq s) p_{jl}^{(s+1)}.$$

Observe $U_i^{(s+1)}$ has a partition consisting of $U_i^{(s)}$ and $O_i^{(s+1)}$, then

$$\sum_{j \in U_i^{(s+1)}} P(Z_j \geq s+1) = \{ \sum_{j \in U_i^{(s)}} \sum_{l \in O_j^{(s+1)}} + \sum_{j \in O_i^{(s+1)}} \sum_{l \in O_j^{(s+1)}} \} P(Z_j \geq s) P(Z_l \geq s) p_{jl}^{(s+1)}$$

Note that $(j, l) \in U_i^{(s)} \times O_j^{(s+1)}$ if and only if $(j, l) \in U_i^{(s)} \times O_i^{(s+1)}$;
$(j, l) \in O_i^{(s+1)} \times O_j^{(s+1)}$ if and only if $(j, l) \in O_i^{(s+1)} \times U_i^{(s)}$, so

$$\begin{aligned}
\sum_{j \in U_i^{(s+1)}} P(Z_j \geq s+1) &= \{ \sum_{j \in U_i^{(s)}} \sum_{l \in O_i^{(s+1)}} + \sum_{j \in O_i^{(s+1)}} \sum_{l \in U_i^{(s)}} \} P(Z_j \geq s) P(Z_l \geq s) p_{jl}^{(s+1)} \\
&= \sum_{j \in O_i^{(s+1)}} \sum_{l \in U_i^{(s)}} P(Z_j \geq s) P(Z_l \geq s) \{ p_{lj}^{(s+1)} + p_{jl}^{(s+1)} \} \\
&= \sum_{j \in O_i^{(s+1)}} \sum_{l \in U_i^{(s)}} P(Z_j \geq s) P(Z_l \geq s) \\
&= \sum_{j \in O_i^{(s+1)}} P(Z_j \geq s) = 1.
\end{aligned}$$

The last equation is due to the fact $O_i^{(s+1)} = U_l^{(s)}$, $\forall\, l \in O_i^{(s+1)}$. □

A.2. R Code for Logistic Conditional Probability Model

```
BB<-read.csv("C:\\Users\\Xiao\\Desktop\\Research\\Data\\NCAA.csv",header=T)
m<- 10*64
X<- as.matrix(BB[1:m,2:6])
```

```
#Rd64#
a<- seq(from=1, to=m-1, by=2)
b<- seq(from=2, to=m, by=2)
Z1<- X[a,]-X[b,];
Y1<- cbind(BB$R1[a],BB$R1[b])
rowSums(Y1)
x1<- Z1[,1]; x2<- Z1[,2]; x3<-Z1[,3]; x4<- Z1[,4]; x5<- Z1[,5]

candi.m9.1<-glm(Y1~0+x1+x2+x5, family=binomial)
m1<-candi.m9.1 #x1,x2,x5#
summary(m1)

plot(resid(m1,type="working")~predict(m1,type="link"),xlab="Link",
ylab="Working Residuals")
######1#########################################################
p<- matrix(NA,nrow=64,ncol=6)
w<- BB[(m+1):704, 2:6]
a1<- seq(from=1, to=63, by=2)
b1<- seq(from=2, to=64, by=2)
z1<- w[a1,c(1,2,5)]-w[b1,c(1,2,5)]
y1<- cbind(BB$R1[a1],BB$R1[b1])
rowSums(y1)
p[a1,1] <- exp(t(coef(m1))%*%t(z1))/(1+exp(t(coef(m1))%*%t(z1)))
p[b1,1] <- exp(t(coef(m1))%*%t(-z1))/(1+exp(t(coef(m1))%*%t(-z1)))

#Rd32#
id<- which(BB$R1==1)
a<- seq(from=1, to=m/2-1, by=2)
b<- seq(from=2, to=m/2, by=2)
```

```
Z2<- X[id[a],]-X[id[b],]
Y2<- cbind(BB$R2[id[a]], BB$R2[id[b]])
rowSums(Y2)
x1<- Z2[,1]; x2<- Z2[,2]; x3<-Z2[,3]; x4<- Z2[,4]; x5<- Z2[,5]

candi.m31.2<-glm(Y2~0+x5, family=binomial)
m2<-candi.m31.2 #x5#
summary(m2)

plot(resid(m2,type="working")~predict(m2,type="link"),xlab="Link",
ylab="Working Residuals")
######2#############################################################
nu<- 2^(6-2)
for (t in 1:nu) {
ri<-2^(2-1); ini<-(t-1)*2*ri; mi<-ini+ri; ui<-ini+2*ri;

for (s in (ini+1):mi) {
psum<- 0
pz<-0
for (i in 1:ri) {
z<- w[s,5]-w[mi+i,5]
pc<- exp(t(coef(m2))%*%t(z))/(1+exp(t(coef(m2))%*%t(z)))
psum<- psum+p[s,1]*p[mi+i,1]*pc
pz<-pz+p[mi+i,1]*pc
                }
p[s,2]<- psum
                 }

for (s in (mi+1): ui) {
```

```
psum<- 0
pz<-0
for (i in 1:ri) {
z<- w[s,5]-w[ini+i,5]
pc<- exp(t(coef(m2))%*%t(z))/(1+exp(t(coef(m2))%*%t(z)))
psum<- psum+p[s,1]*p[ini+i,1]*pc
                }
p[s,2]<- psum
                      }
       }

#Sweet16#
id<- which(BB$R2==1)
a<- seq(from=1, to=m/4-1, by=2)
b<- seq(from=2, to=m/4, by=2)
Z3<- X[id[a],]-X[id[b],]
Y3<- cbind(BB$R3[id[a]], BB$R3[id[b]])
rowSums(Y3)
x1<- Z3[,1]; x2<- Z3[,2]; x3<-Z3[,3]; x4<- Z3[,4];x5<- Z3[,5]

candi.m31.3<-glm(Y3~0+x5, family=binomial)
m3<-candi.m31.3 #x5#
summary(m3)
######3#############################################
nu<- 2^(5-2)
for (t in 1:nu) {
ri<-2^(3-1); ini<-(t-1)*2*ri; mi<-ini+ri; ui<-ini+2*ri;

for (s in (ini+1):mi) {
```

```
psum<- 0
for (i in 1:ri) {
z<- w[s,5]-w[mi+i,5]
pc<- exp(t(coef(m3))%*%t(z))/(1+exp(t(coef(m3))%*%t(z)))
psum<- psum+p[s,2]*p[mi+i,2]*pc
                }
p[s,3]<- psum
                 }

for (s in (mi+1): ui) {
psum<- 0
for (i in 1:ri) {
z<- w[s,5]-w[ini+i,5]
pc<- exp(t(coef(m3))%*%t(z))/(1+exp(t(coef(m3))%*%t(z)))
psum<- psum+p[s,2]*p[ini+i,2]*pc
                }
p[s,3]<- psum
                      }
       }

#Elite8#
id<- which(BB$R3==1)
a<- seq(from=1, to=m/8-1, by=2)
b<- seq(from=2, to=m/8, by=2)
Z4<- X[id[a],]-X[id[b],]
Y4<- cbind(BB$R4[id[a]], BB$R4[id[b]])
rowSums(Y4)
x1<- Z4[,1]; x2<- Z4[,2]; x3<-Z4[,3]; x4<- Z4[,4];x5<- Z4[,5]
```

```
candi.m31.4<-glm(Y4~0+x5, family=binomial)
m4<-candi.m31.4 #x5#
summary(m4)

plot(resid(m4,type="working")~predict(m4,type="link"),xlab="Link",
ylab="Working Residuals")
######4################################################
nu<- 2^(4-2)
for (t in 1:nu) {
ri<-2^(4-1); ini<-(t-1)*2*ri; mi<-ini+ri; ui<-ini+2*ri;

for (s in (ini+1):mi) {
psum<- 0
for (i in 1:ri) {
z<- w[s,5]-w[mi+i,5]
pc<- exp(t(coef(m4))%*%t(z))/(1+exp(t(coef(m4))%*%t(z)))
psum<- psum+p[s,3]*p[mi+i,3]*pc
                }
p[s,4]<- psum
                 }

for (s in (mi+1): ui) {
psum<- 0
for (i in 1:ri) {
z<- w[s,5]-w[ini+i,5]
pc<- exp(t(coef(m4))%*%t(z))/(1+exp(t(coef(m4))%*%t(z)))
psum<- psum+p[s,3]*p[ini+i,3]*pc
                }
p[s,4]<- psum
```

```
                    }
        }

#Final4#
id<- which(BB$R4==1)
a<- seq(from=1, to=m/16-1, by=2)
b<- seq(from=2, to=m/16, by=2)
Z5<- X[id[a],]-X[id[b],]
Y5<- cbind(BB$R5[id[a]], BB$R5[id[b]])
rowSums(Y5)
x1<- Z5[,1]; x2<- Z5[,2]; x3<-Z5[,3]; x4<- Z5[,4];x5<- Z5[,5]

candi.m26.5<-glm(Y5~0+x4+x5, family=binomial)
m5<-candi.m26.5 #x4,x5#
summary(m5)

plot(resid(m5,type="working")~predict(m5,type="link"),xlab="Link",
ylab="Working Residuals")
######5################################################
nu<- 2^(3-2)
for (t in 1:nu) {
ri<-2^(5-1); ini<-(t-1)*2*ri; mi<-ini+ri; ui<-ini+2*ri;

for (s in (ini+1):mi) {
psum<- 0
for (i in 1:ri) {
z<- w[s,4:5]-w[mi+i,4:5]
pc<- exp(t(coef(m5))%*%t(z))/(1+exp(t(coef(m5))%*%t(z)))
psum<- psum+p[s,4]*p[mi+i,4]*pc
```

```
                }
p[s,5]<- psum
                 }

for (s in (mi+1): ui) {
psum<- 0
for (i in 1:ri) {
z<- w[s,4:5]-w[ini+i,4:5]
pc<- exp(t(coef(m5))%*%t(z))/(1+exp(t(coef(m5))%*%t(z)))
psum<- psum+p[s,4]*p[ini+i,4]*pc
                }
p[s,5]<- psum
                      }
       }

#Championship#
id<- which(BB$R5==1)
a<- seq(from=1, to=m/32-1, by=2)
b<- seq(from=2, to=m/32, by=2)
Z6<- X[id[a],]-X[id[b],]
Y6<- cbind(BB$R6[id[a]], BB$R6[id[b]])
rowSums(Y6)
x1<- Z6[,1]; x2<- Z6[,2]; x3<-Z6[,3]; x4<- Z6[,4];x5<- Z6[,5]

candi.m31.6<-glm(Y6~0+x5, family=binomial)
m6<-candi.m31.6 #x5#
summary(m6)
######6################################################
nu<- 2^(2-2)
```

```
for (t in 1:nu) {
ri<-2^(6-1); ini<-(t-1)*2*ri; mi<-ini+ri; ui<-ini+2*ri;

for (s in (ini+1):mi) {
psum<- 0
for (i in 1:ri) {
z<- w[s,5]-w[mi+i,5]
pc<- exp(t(coef(m6))%*%t(z))/(1+exp(t(coef(m6))%*%t(z)))
psum<- psum+p[s,5]*p[mi+i,5]*pc
                }
p[s,6]<- psum
                 }

for (s in (mi+1): ui) {
psum<- 0
for (i in 1:ri) {
z<- w[s,5]-w[ini+i,5]
pc<- exp(t(coef(m6))%*%t(z))/(1+exp(t(coef(m6))%*%t(z)))
psum<- psum+p[s,5]*p[ini+i,5]*pc
                }
p[s,6]<- psum
                      }
       }

#expected wins for each team in 2012-2013 season#
rowSums(p)

#Actual wins for each team in 2012-2013 season#
ExpWins<-c(6,0,1,0,0,2,1,0,1,0,2,0,1,0,3,0,
```

```
2,0,1,0,1,0,5,0,0,1,3,0,1,0,0,2,
2,0,0,1,0,1,4,0,1,0,3,0,1,0,2,0,
1,0,0,4,0,1,0,2,2,0,0,1,0,1,3,0)

###read probs and expected wins for 2012-2013 season as a table###
pe<-cbind(p,rowSums(p),ExpWins)

setwd("C:\\Users\\Xiao\\Desktop\\Research\\Data")
getwd()
write.table(pe,"Shen's.csv",sep=",")
```

A.3. SAS and R Code for Restricted OLRE Model

A.3.1. SAS Code Part

```
data NCAA_OLRE;
  infile 'C:\\Users\\Xiao\\Desktop\\Research\\Data\\NCAA_OLRE.csv' dlm=','
  firstobs=2 missover DSD;
  input Wins DIFF SAGSOS ATRATIO APG Pyth;
  cards;
  run;
proc logistic data = NCAA_OLRE;
   model Wins = DIFF SAGSOS ATRATIO APG Pyth;
   output out = newdata predprobs = individual;
run;
proc print data=newdata;
run;
```

A.3.2. R Code Part

```
source("mph.r")
source("mph.supp.fcts.r")
```

```
y <- scan()
0.0271  0.1024  0.1949  0.2292  0.1871  0.1211  0.1382
0.9996  0.0003  0.0001  0.0000  0.0000  0.0000  0.0000
0.3080  0.3960  0.1808  0.0672  0.0266  0.0115  0.0099
0.2469  0.3897  0.2132  0.0862  0.0352  0.0154  0.0134
0.2292  0.3845  0.2231  0.0930  0.0385  0.0169  0.0148
0.3973  0.3816  0.1403  0.0479  0.0184  0.0078  0.0067
0.2421  0.3884  0.2158  0.0879  0.0360  0.0158  0.0138
0.9214  0.0629  0.0108  0.0030  0.0011  0.0005  0.0004
0.5039  0.3405  0.1016  0.0324  0.0121  0.0051  0.0044
0.2763  0.3947  0.1972  0.0763  0.0306  0.0133  0.0116
0.1312  0.3153  0.2760  0.1480  0.0688  0.0319  0.0287
0.7958  0.1584  0.0312  0.0089  0.0032  0.0014  0.0011
0.2174  0.3801  0.2299  0.0979  0.0409  0.0181  0.0158
0.3538  0.3914  0.1590  0.0564  0.0219  0.0094  0.0081
0.0648  0.2054  0.2743  0.2109  0.1214  0.0628  0.0605
0.9911  0.0072  0.0012  0.0003  0.0001  0.0000  0.0000
0.0780  0.2332  0.2821  0.1970  0.1064  0.0532  0.0501
0.9971  0.0023  0.0004  0.0001  0.0000  0.0000  0.0000
0.3983  0.3813  0.1399  0.0477  0.0183  0.0078  0.0067
0.6189  0.2778  0.0689  0.0208  0.0077  0.0032  0.0027
0.1676  0.3506  0.2582  0.1233  0.0541  0.0245  0.0217
0.8009  0.1547  0.0303  0.0087  0.0031  0.0013  0.0011
0.0712  0.2193  0.2788  0.2041  0.1137  0.0578  0.0550
0.9515  0.0390  0.0065  0.0018  0.0006  0.0003  0.0002
0.5696  0.3065  0.0819  0.0253  0.0094  0.0039  0.0034
0.2919  0.3958  0.1890  0.0717  0.0285  0.0124  0.0107
0.0255  0.0973  0.1885  0.2273  0.1903  0.1255  0.1455
```

0.9904	0.0078	0.0013	0.0003	0.0001	0.0001	0.0000
0.3980	0.3814	0.1400	0.0478	0.0183	0.0078	0.0067
0.5908	0.2944	0.0762	0.0233	0.0086	0.0036	0.0031
0.2140	0.3786	0.2318	0.0994	0.0416	0.0184	0.0161
0.9603	0.0320	0.0053	0.0015	0.0005	0.0002	0.0002
0.0263	0.0998	0.1917	0.2283	0.1887	0.1233	0.1418
0.9964	0.0029	0.0005	0.0001	0.0000	0.0000	0.0000
0.4045	0.3795	0.1374	0.0466	0.0178	0.0076	0.0065
0.7562	0.1869	0.0386	0.0112	0.0041	0.0017	0.0014
0.5407	0.3221	0.0903	0.0282	0.0105	0.0044	0.0038
0.7465	0.1937	0.0404	0.0117	0.0043	0.0018	0.0015
0.0823	0.2417	0.2834	0.1924	0.1021	0.0506	0.0474
0.9949	0.0041	0.0007	0.0002	0.0001	0.0000	0.0000
0.5682	0.3073	0.0823	0.0254	0.0094	0.0040	0.0034
0.7224	0.2105	0.0453	0.0132	0.0048	0.0020	0.0017
0.3862	0.3845	0.1449	0.0499	0.0192	0.0082	0.0070
0.6803	0.2388	0.0543	0.0161	0.0059	0.0025	0.0021
0.4412	0.3672	0.1232	0.0408	0.0155	0.0066	0.0056
0.5359	0.3246	0.0917	0.0287	0.0107	0.0045	0.0039
0.1432	0.3285	0.2707	0.1392	0.0633	0.0291	0.0260
0.9909	0.0074	0.0012	0.0003	0.0001	0.0000	0.0000
0.0525	0.1759	0.2603	0.2230	0.1387	0.0750	0.0745
0.9934	0.0054	0.0009	0.0002	0.0001	0.0000	0.0000
0.1068	0.2830	0.2836	0.1685	0.0827	0.0394	0.0360
0.3156	0.3957	0.1770	0.0653	0.0257	0.0111	0.0096
0.0844	0.2457	0.2839	0.1902	0.1002	0.0494	0.0462
0.2825	0.3953	0.1939	0.0744	0.0298	0.0129	0.0112
0.3899	0.3836	0.1434	0.0493	0.0189	0.0081	0.0069
0.6484	0.2594	0.0617	0.0185	0.0068	0.0028	0.0024

```
0.2085  0.3761  0.2350  0.1019  0.0428  0.0190  0.0167
0.5594  0.3121  0.0848  0.0263  0.0098  0.0041  0.0035
0.3424  0.3932  0.1642  0.0589  0.0230  0.0099  0.0085
0.9621  0.0305  0.0050  0.0014  0.0005  0.0002  0.0002
0.4366  0.3688  0.1249  0.0415  0.0157  0.0067  0.0057
0.3730  0.3877  0.1505  0.0524  0.0202  0.0087  0.0074
0.1540  0.3390  0.2654  0.1318  0.0590  0.0269  0.0239
0.9536  0.0374  0.0062  0.0017  0.0006  0.0003  0.0002

Z <- ZF <- pop(64, 7)

h.fct <- function(m) {
   p <- diag(c(1/Z%*%t(Z)%*%m))%*%m
   p <- matrix(p, 64, 7, byrow=T)
   colsum <- apply(p, 2, sum)[-7]
   as.matrix(colsum -c(32, 16, 8, 4, 2, 1))
}

a <- mph.fit(y, Z, ZF, h.fct=h.fct, norm.diff.conv=10, norm.score.conv=1e-8)
y.matrix <- matrix(y, 64, 7, byrow=T)
fitted.value.matrix <- matrix(a$m, 64, 7, byrow=T)
apply(y.matrix, 2, sum)
apply(fitted.value.matrix, 2, sum)
fitted.value.matrix

setwd("C:\\Users\\Xiao\\Desktop\\Research\\Data")
getwd()
write.table(fitted.value.matrix, "adjusted_probs_0413.csv", sep=",")
```