PREDICTING THE OUTCOMES OF NCAA WOMEN'S SPORTS

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

By

Wenting Wang

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Statistics

November 2017

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Predicting the Outcomes of NCAA Women's Sports

By

Wenting Wang

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

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Rhonda Magel
Chair

Dr. Gang Shen

Dr. Ronald Degges

Dr. Zhaohui Liu

Approved:

11/16/2017
Date

Rhonda Magel
Department Chair

ABSTRACT

Sports competitions provide excellent opportunities for model building and using basic statistical methodology in an interesting way. More attention has been paid to and more research has been conducted pertaining to men's sports as opposed to women's sports. This paper will focus on three kinds of women's sports, i.e. NCAA women's basketball, volleyball and soccer.

Several ordinary least squares models were developed that help explain the variation in point spread of a women's basketball game, volleyball game and soccer game based on in-game statistics. Several logistic models were also developed that help estimate the probability that a particular team will win the game for women's basketball, volleyball and soccer tournaments.

Ordinary least squares models for Round 1, Round 2 and Rounds 3-6 with point spread being the dependent variable by using differences in ranks of seasonal averages and differences of seasonal averages were developed to predict winners of games in each of those rounds for the women's basketball, volleyball and soccer tournament. Logistic models for Round 1, Round 2 and Rounds 3-6 that estimate the probability of a team winning the game by using differences in ranks of seasonal averages and differences of seasonal averages were developed to predict winners of games in each of those rounds for the basketball, volleyball and soccer tournaments.

The prediction models were validated before doing the prediction. For basketball, the least squares model developed by using differences in ranks of seasonal averages with a double scoring system variable predicted the results of a 76.2% of the games for the entire tournament with all the predictions made before the start of the tournament. For volleyball, the logistic model developed by using differences of seasonal averages predicted 65.1% of the games for the entire tournament. For soccer, the logistic regression model developed by using differences of seasonal averages predicted 45% of all games in the tournament. Correctly when all 6 rounds

were predicted before the tournament began. In this case, team predicted to win in the second round or higher might not have even made it to this round since prediction was done ahead of time.

ACKNOWLEDGMENTS

It would not have been possible to write this doctoral thesis without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

It is a great pleasure to acknowledge my deepest thanks and gratitude to my Supervisor, Dr. Rhonda Magel for her meaningful assistance, tireless guidance and patience over the years which is unmeasurable and without it I would not be where I am today. I thank her so much for the knowledge she has passed on and I will always be grateful for having the opportunity to study under her.

I would also like to acknowledge helpful suggestions from my committee members: Dr. Gang Shen, Dr. Ronald Degges, and Dr. Zhaohui Liu. Thanks for their endless help, generous advice and support during the study. This work would not have been possible without their help and input.

I am also grateful to Sandie Salisbury, my advisor at Noridian Healthcare Solutions, for providing me flexible working hours to complete writing this paper. It was particularly kind of her to allow me to take days off to finish my writing. It is a great honor to work under her supervision.

Finally, I would like to thank my husband Guojia Ma for his personal support and great patience at all times. Big thanks to my two children, Oscar and Alvin, they always been a constant source of joy when I am facing struggles and difficulties, they are always the motivation for me to complete my degree. I also want to thank my parents and parents in laws, they have given me their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

The use of statistics in sports has drawn more and more interest in the past several years. Female participation and popularity in sports also increased dramatically in recent years (Women's sports [1]). Women's basketball, volleyball and soccer are the most well-known and competitive sports with physical and technical performances over the years.

The NCAA division I women's basketball tournament is an annual college basketball tournament for women. It is also known as March Madness or The Big Dance since it is staged in a single elimination format. Unlike the men's tournament, there are only 32 at-large bids and no play-in games (NCAA – Basketball [2]).

The NCAA division I women's volleyball championship is the annual championship in women's volleyball from teams in division I contested by the NCAA each winter since 1981.Volleyball was one of twelve women's sports added to the NCAA championship program for the 1981-1982 school year (NCAA – Volleyball [3]).

The NCAA division I women's soccer championship is also known as the women's College Cup. It is an American college soccer tournament conducted by National Collegiate Athletic Association (NCAA) (NCAA – Soccer [4]).

This research will focus on developing models that help explain the point spread between the two teams participating in an NCAA women's game of basketball, volleyball, and soccer. These models will be used and additional models will be developed to help predict the outcomes of NCAA tournaments in these sports.

1.1. References

[1] Women's sports. Retrieved October 20, 2017, from https://en.wikipedia.org/wiki/Women%27s_sports

[2] NCAA Division I Women's Basketball Tournament. Retrieved October 10, 2017, from https://en.wikipedia.org/wiki/NCAA_Division_I_Women%27s_Basketball_Tournament

[3] NCAA Division I Women's Volleyball Championship. Retrieved October 10, 2017, from https://en.wikipedia.org/wiki/NCAA_Division_I_Women%27s_Volleyball_Championship

[4] NCAA Division I Women's Soccer Championship. Retrieved October 10, 2017, from https://en.wikipedia.org/wiki/NCAA_Division_I_Women%27s_Soccer_Championship

CHAPTER 2. REVIEW OF PAST STUDIES

More attention has been paid to and more research has been conducted pertaining to men's sports as opposed to women's sports. Some of the findings will be presented.

2.1. Basketball

Schwertman et al. (1996) [1] developed models to estimate the probability of any given men's basketball team winning their regional tournament advancing to the 'Final Four'. They modified the approach to fit linear and logistic regression models as a function of the difference in either team seeds or normal scores of the team seeds. One variable that was placed into their models was the team's overall seed in the tournament.

Kubatko et al. (2007) [2] introduced some basic basketball statistics to consider when analyzing men's basketball games. They found in-game statistics are useful in the diagnostics of the performance of a team and helpful for the team to prepare a future since the in-game statistics measures a different dimension of the performance of a team in a game.

Magel and Unruh (2013) [3] collected statistics from two seasons of NCAA men's basketball games and used these to develop logistic and ordinary least squares regression models. Difference in assists, free throw attempts, defensive rebounds and turnovers were found to be significant to determining victory. The models were verified by using the data from 2011-2012 season and used in prediction for 2013 NCAA tournament.

Shen et al. (2015) [4] developed a new bracketing tool for the NCAA men's basketball tournament. The method is based on a binomial generalized linear regression model with Cauchy link on the conditional probability of a team winning a game given its rival team. The new method then was compared to three existing methods to help complete March Madness brackets and the result shows their new method did better than the other three methods in predicting March Madness winners.

Jones and Magel (2016) [5] developed models based on a stratified random sample of 144 NBA basketball games. Field goal shooting percentage, three-point shooting percentage, free throw shooting percentage, offensive rebounds, assists, turnovers and free throws attempted were found to be significant when developing the models. The models were validated using a random sample of 50 NBA games and then were used to do the predictions.

Huang and Magel (2016) [6] developed ordinary least squares regression models and logistic regression models of NCAA women's division II basketball tournament game using in-game statistics. The models were verified based on a random sample of basketball games and then used to predict the outcomes of the 2015 NCAA dvision II women's basketball tournament.

2.2. Volleyball

Giatsis (2008) [7] conducted an analysis on men's beach volleyball. The purpose of his study was to determine the differences in playing characteristics between winning and losing teams in FIVB Men's Beach Volleyball World Tour Tournament. Giatsis used independent t-tests and a discriminant function analysis to determine which skills contributed significantly to winning in matches. He found the opponents' attack errors was the most significant factor contributing to winner's win.

Zhang (2016) [8] developed a multiple linear regression model using in-game statistics that explain the point spread of a volleyball game and a logistic regression model that estimates the probability of a team winning the game based on the in-game statistics for women's volleyball game. The point spread model was used to predict the results of future volleyball game by replacing the in-game statistics with the averages of the in-game statistics based on the past two previous matches of both teams playing each other.

2.3. Soccer

Parentos (2012) [9] developed a model to determine factors in the men's European soccer Champions League that influence the number of goals that a team scores. Ball possession percentage and the logarithm of the ratio between goals scored and goals conceded were the two of the variables that he considered.

Magel and Melnykov (2014) [10] conducted an analysis of games played by three top men's European soccer leagues during the first 33 rounds of soccer during the 2011-2012 season. They developed two regression models to predict the point spread of a game between two teams. The models correctly predicted the winner of a game at 73% to 80% of the time when predict winners of games for the last five rounds of the 2011-2012 season.

Sylla and Magel (2016) [11] developed several statistical models to predict the outcomes of men's World Cup soccer matches. Data from the 2006 World Cup Matches was used to develop the ordinary least squares regression and logistic regression models. These models were then tested by using data from the 2010 World Cup Matches and then used to predicted the 2014 World Cup Championship.

2.4. Football

Willoughby (2002) [12] conducted an analysis of games on the men's Canadian Football league by using logistic regression in order to determine factors leading to a team's overall success. Willoughby separated teams into three categories: 'very good' teams, 'average' teams, and 'poor' teams. The models were developed by using in-game statistics and found the models predicted 85.9% correctly for 'very good' teams, 90.2% correctly for 'average' teams and 78.8 correctly for 'poor' teams.

Karlis and Ntzoufras (2003) [13] used an indirect approach and modeled the goals scored by each team playing in the match using a bivariate Poisson model. They found using a bivariate Poisson distribution can improve model fit and prediction of the number of draws in men's football games.

Magel and Long (2013) [14] conducted an analysis of games played by FCS Division I men's college football. They developed models by using in-game statistics to estimate point spread of the game and the probability that a particular team will win the football game when the in-game statistics are known. These models are then used to predict the outcome of future football games.

2.5. Description of Study

In this paper, we will focus on three kinds of women's sports, i.e. women's basketball, women's volleyball, and women's soccer. First, we would like to develop several models that help explain the variation in point spread of a NCAA women's division I basketball game, women's volleyball game and women's soccer game based on in-game statistics. We would also like to develop several logistic regression models that help estimate the probability that a particular team will win the game. Various sets of statistics will be used to develop the models for each sport. Once the models are developed they will be validated.

After developing models to explain point spread of games using in-game statistics and also developing models to estimate the probability of a team winning based on given differences of in-game statistics, we would like to develop prediction models. The prediction models will be based on each round of a NCAA tournament game using various seasonal statistics. These statistics are often differences of seasonal averages between the two teams playing or differences in ranks of seasonal averages of the two teams playing. Results will be given.

2.6. References

[1] Schertman, N.C., Schenk, K.L., and Holdbrook, B.C., (1996). *More probability Models for the NCAA Regional Basketball Tournaments*. The American Statistician, 50: 34-38

[2] Kubatko, J. and Olicer, 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

[3] R. Magel, S. Unruh (2013). *Determining Factor Influencing the Outcome of College Basketball Games*, Open Journal of Statistics, Vol.3 No. 4, 2013, p. 225-230

[4] G. Shen, et al. (2015). *Predicting Results of March Madness Using the Probability Self-Consistent Method*, International Journal of Sports Science, Vol. 5(4), 139-144

[5] E. Jones, R. Magel (2016). *Predicting Outcomes of NBA Basketball Games, Journal of Advance Research in Business*, Management and Accounting, Vol. 3, Issue 5

[6] F. Huang, R. Magel (2016). *Developing Models to Explain Point Spread of NCAA Women's Division II Basketball Games*, JIATTS, June 2016

[7] Giatsis, George (2008). *Statistical Analysis of Men's FIVB Beach Volleyball Team Performance*. International Journal of Performance Analysis in Sport, 31-43

[8] D. Zhang, (2016). *Forecasting Point Spread for Women's Volleyball*. Unpublished Thesis Paper, North Dakota State University, Fargo, ND

[9] V. Panaretos (2012). *A statistical analysis of the European Soccer Champions League*, Joint Statistical Meetings – Section on Statistics in Sports, 2600-2602

[10] R. Magel, Y. Melnykov (2014), *Examining Influential Factors and Predicting Outcomes in European Soccer Games*, International Journal of Sports Science, Vol. 4, No. 3

[11] M. Sylla, R. Magel (2016). *Predicting the Winner of Games in World Cup Soccer Matches*, Advance Research in Mathematics and Mathematical Sciences, Vol. 1, Issue 8

[12] Willoughby, K. A. (2002). *Winning Games in Canadian Football: A Logistic Regression Analysis*. The College Mathematics Journal 33(3):215-220

[13] Karlis, D., Ntzoufras, J. (2003). *Analysis of sports data using bivariate Poisson models*. The Statistician,52, 381-393

[14] R. Magel, J. Long (2013). *Identifying Significant In-Game Statistics and Developing Prediction Models for Outcomes of NCAA Division I Football Championship Subdivision(FCS) Games*, Journal of Statistical Science and Application, Vol. 1, No., 51-62

CHAPTER 3. BRACKETING NCAA WOMEN'S BASKETBALL TOURNAMENT

3.1. Introduction

3.1.1. The history of NCAA women's basketball tournament

The NCAA division I women's basketball tournament is an annual championship in women's basketball from teams in division I contested by the National Collegiate Athletic Association(NCAA) each spring since 1982. Basketball was one of the 12 women's sports added to the NCAA championship program for the 1981-1982 school year. There were only 32 teams competing for the first NCAA championship which was held in 1982. The tournament expanded gradually, and reached its current size of 64 teams in 1994 (NCAA – Basketball [1]).

3.1.2. The playing rule and structure

The significant difference between the women's and men's basketball tournament is that in the women's tournament, there are still only 64 teams with 32 at-large bids and no play-in games as in the men's tournament. For the tournament bracket, champions from each division I conference receive automatic bids. The remaining slots are at-large bids, with teams chosen by an NCAA selection committee. The selection process is based on team rankings, win-loss records and Ratings Percentage Index (RPI) data (NCAA – Basketball [1]). Like the men's tournament, the women's tournament is staged in a single elimination format, this made the games very watchable and is part of the reason why it is also known as March Madness or The Big Dance (Road to the Championship [2]).

For the first round, there will be 64 teams compete in single-elimination for second round. The 32 advancing teams then compete against each other in single-elimination second round competition. The winning team will advance to the third round. For the third round, there will be 16 teams compete in single-elimination regional semifinal competition. The advancing teams then compete against each other in single-elimination regional final. The winning team in each of the four regions will advance to the NCAA women's basketball semifinal round. There will be 4 teams competing in the single-elimination semifinal and the advancing teams then compete against each other for the national championship title (Road to the Championship [2]). Figure 1 shows the 2015 - 2016 NCAA women's basketball tournament bracket.

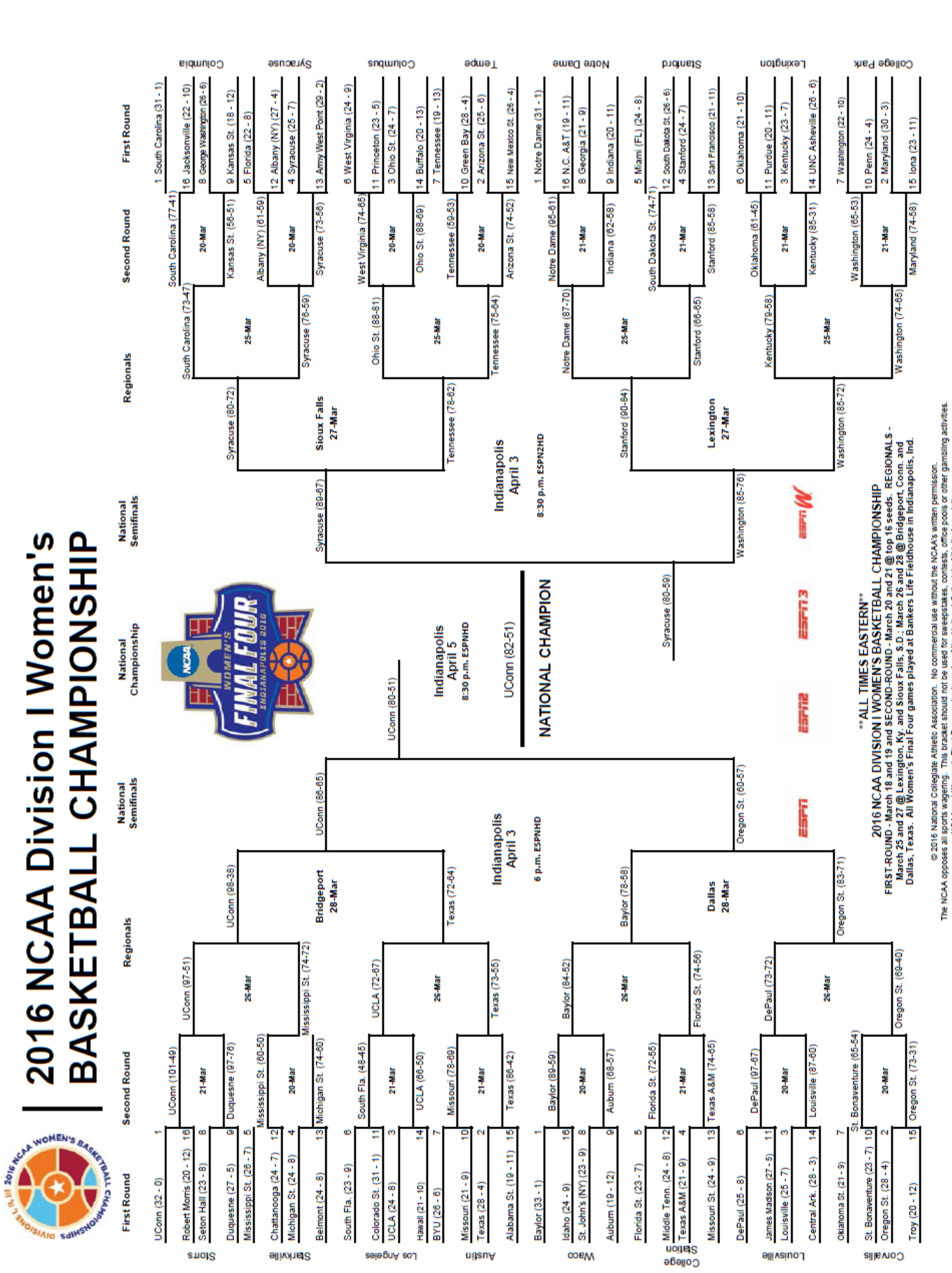

Figure 1. The NCAA women's basketball tournament bracket for the 2015 – 2016 season. (This bracket is downloaded from: http://www.ncaa.com/interactive-bracket/basketball-women/d1)

3.1.3. The research objectives for this study

The research objectives for this study are as follows:

1) Develop ordinary least squares regression models with point spread as dependent variables for Round 1, Round 2 and Rounds 3-6 by using differences in ranks of seasonal averages with a single scoring system variable, to predict winners of basketball games in each of those rounds for the NCAA women's basketball tournament; and

2) Develop ordinary least squares regression models with point spread as dependent variables for Round 1, Round 2 and Rounds 3-6 by using differences in ranks of seasonal averages with a double scoring system variable, to predict winners of basketball games in each of those rounds for the NCAA women's basketball tournament; and

3) Develop logistic regression models for Round 1, Round 2 and Rounds 3-6 that estimate the probability of a team winning the game by using differences in ranks of seasonal averages with a single scoring system variable, to predict winners of basketball games in each of those rounds for the NCAA women's basketball tournament; and

4) Develop logistic regression models for Round 1, Round 2 and Rounds 3-6 that estimate the probability of a team winning the game by using differences in ranks of seasonal averages with a double scoring system variable, to predict winners of basketball games in each of those rounds for the NCAA women's basketball tournament; and

5) Develop ordinary least squares regression models with point spread as dependent variables for Round 1, Round 2 and Rounds 3-6 by using differences of seasonal averages with a single scoring system variable, to predict winners of basketball games in each of those rounds for the NCAA women's basketball tournament; and

6) Develop ordinary least squares regression models with point spread as dependent variables for Round 1, Round 2 and Rounds 3-6 by using differences of seasonal averages with a double scoring system variable, to predict winners of basketball games in each of those rounds for the NCAA women's basketball tournament; and

7) Develop logistic regression models for Round 1, Round 2 and Rounds 3-6 that estimate the probability of a team winning the game by using differences of seasonal averages with a single scoring system variable, to predict winners of basketball games in each of those rounds for the NCAA women's basketball tournament; and

8) Develop logistic regression models for Round 1, Round 2 and Rounds 3-6 that estimate the probability of a team winning the game by using differences of seasonal averages with a double scoring system variable, to predict winners of basketball games in each of those rounds for the NCAA women's basketball tournament; and

9) Develop one ordinary least squares regression model with point spread as dependent variable using in-game statistics, to explain the variation in point spread of basketball games for the NCAA women's basketball tournament; and

10) Develop one logistic regression model for round 1-6 that estimate the probability of a team winning the game by using in-game statistics, to predict winners of basketball games for the NCAA women's basketball tournament.

In order to accomplish objectives 1 to 8, data was collected for three years of the NCAA women's basketball tournament. This included 2011, 2012 and 2013 tournaments. Seasonal averages and the ranks of the seasonal averages were collected for all the teams in the 2011 tournament on the variables listed in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-

Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin and Three Pt FG Defense. Seasonal averages were also collected on the same variables for all teams playing in the 2012 and 2013 tournaments. Seasonal average statistics from each of the teams were collected from the official NCAA Basketball statistics Database (NCAA [3]). Data was collected before the tournament started. For example, the first game of NCAA 2011 women's basketball tournament was held on March 19, 2011. The seasonal averages for each of the variables were based on games played through March 16, 2011.

Table 3.1. Set A - Variables in consideration for seasonal averages

Variables in consideration	Definitions
Scoring Offense	SO = $\frac{Total\ points}{Total\ number\ of\ games\ played}$ [4]
Scoring Defense	SD = $\frac{Total\ points\ made\ by\ opponent}{Total\ number\ of\ games\ played}$ [4]
Scoring Margin	SM = Scoring Offense – Scoring Defense [4]
Field-Goal Percentage	FG% = $\frac{Field\ goal\ made}{Field\ goal\ attempt}$ [4]
Field-Goal Percentage Defense	FGPD = $\frac{Field\ goal\ made\ by\ opponent}{Field\ goal\ attempt\ by\ opponent}$ [4]
Free-Throw Percentage	FT% = $\frac{Free\ throw\ made}{Free\ throw\ attempt}$ [4]
Rebound Margin	RM = $\frac{\text{Number of Rebounds}}{Total\ game\ played}$ - $\frac{Number\ of\ Rebounds\ by\ opponent}{Total\ game\ played\ by\ opponent}$ [4]
Three-Point Field Goals Per Game	TPFGPG = $\frac{Total\ three\ point\ field\ goals\ made}{Total\ number\ of\ games\ played}$ [4]
Three-Point Field Goal Percentage	TPFGP = $\frac{Total\ three\ point\ field\ goals\ made}{Total\ three\ point\ field\ goals\ attempt}$ [4]
Won-Lost Percentage	WLP = $\frac{Total\ game\ won}{Total\ game\ played}$ [4]
Assists Per Game	APG = $\frac{Total\ number\ of\ assists}{Total\ games\ played}$ [4]
Blocked Shots Per Game	BSPG = $\frac{Total\ number\ of\ blocked\ shots}{Total\ games\ played}$ [4]
Steals Per Game	SPG = $\frac{Total\ number\ of\ steals}{Total\ games\ played}$ [4]
Turnovers Per game	TPG = $\frac{Total\ number\ of\ turnovers}{Total\ games\ played}$ [4]
Personal Fouls Per Game	PFPG = $\frac{Total\ number\ of\ fouls}{Total\ games\ played}$ [4]
Assist Turnover Ratio	ATR = $\frac{\text{Number of assists made}}{Total\ game\ played}$ / $\frac{Number\ of\ turnover\ made}{Total\ game\ played}$ [4]
Turnover Margin	TM = $\frac{Number\ of\ turnover\ by\ opponent}{Total\ game\ played\ by\ opponent} - \frac{\text{Number of turnovers}}{Total\ game\ played}$ [4]
Three Pt FG Defense	TPFGD = $\frac{Three\ point\ field\ goals\ made\ by\ opponent}{Three\ point\ field\ goals\ attempt\ by\ opponent}$ [4]

For research objectives 9 and 10, data was collected for NCAA women's basketball tournament of 2014. In-game statistics were collected for 63 games of the 2014 tournament on the variables listed in Table 3.2 (Set B). The variables included: Free-Throw Percentage (FT%), Field-Goal Percentage (FG%), 3 Point Goals Percentage (3P%), Offensive Rebounds (OREB), Assists (AST), Steals (ST), Blocks (BLK) and Turnovers (TO).

Table 3.2. Set B - Variables in consideration for in-game statistics

Variables in consideration	**Definitions**
Free-Throw Percentage (FT%)	An unguarded shot taken from the foul line by a player whose opponent committed a personal or technical foul; it is worth 1 point. FT%=$\frac{Free\ throw\ made}{Free\ throw\ attempt}$ [5] [6]
Field-Goal Percentage (FG%)	A basket scored on any shot other than a free throw, worth two or three points depending on the distance of the attempt from the basket. FG%=$\frac{Field\ goal\ made}{Field\ goal\ attempt}$ [5] [6]
3 Point Goals Percentage (3P%)	A field goal worth 3 points because the shooter had both feet behind the 3-point line when he released the ball. 3P%=$\frac{3\ point\ goals\ made}{3\ point\ goals\ attempt}$ [5] [6]
Offensive Rebounds (OREB)	A rebound by a player on offense. [5]
Assists (AST)	The last pass to a teammate that leads directly to a field goal; the scorer must move immediately toward the basket for the passer to be credited with an assist; only 1 assist can be credited per field goal. [5] [6]
Steals (ST)	To take the ball away from the opposing team, either off the dribble or by picking off a pass. [5]
Blocks (BLK)	The successful deflection of a shot by touching part of the ball on its way to the basket, thereby preventing a field goal. [5]
Turnovers (TO)	When the offense loses possession through its own fault by passing the ball out of bounds or committing a floor violation. [5]

3.2. Develop models by using differences in ranks of seasonal averages

Ranks of the seasonal averages from each of the team were collected and differences of these ranks were found from the official NCAA basketball statistics database (NCAA [3]). Data was collected for three years of the women's basketball tournament and collected before the tournament started. For example, the first game of NCAA 2011 women's basketball tournament began on March 19, 2011, the seasonal averages and their ranks were based on all games played through March 16, 2011. Seasonal averages were also collected before the 2012 and 2013 tournaments began, ranks of these seasonal averages were found, and difference taken on the variables listed in Table 3.1 (Set A).

3.2.1. Bracket scoring system

Variables representing points received by teams for the previous two years based on the single scoring system and the double scoring system (Shen et al., 2015 [7]) of March Madness were separately considered for entry into the model.

In the single scoring system, a team will be rewarded one single point for each game they win in the March Madness Tournament. There are 6 rounds in the tournament, so 6 is the maximum number of points a team could receive in one tournament of March Madness.

For double scoring system, a team will receive one point for winning the first round, and 2 points for winning the second round and the points a team would receive are for winning a game doubled for each consecutive round in March Madness. The maximum points will be 63 for one tournament. This gives increasingly more weight to games won as the tournament unfolds, presumably to reflect the increasing importance of each round.

Table 3.3 gives the number of points a team would receive for winning the round under each scoring system.

Table 3.3. Single scoring system and double scoring system

Variable	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Max
Single scoring	1	1	1	1	1	1	6
Double scoring	1	2	4	8	16	32	63

For each team playing in the 2011, 2012, and 2013 March Madness Tournaments, the number of points under the single scoring system, and then under the double scoring system for the two previous years in the tournament was found.

As an example, Connecticut played in the tournament in 2011. The points Connecticut received under the single and double scoring system for the previous two years are calculated in Tables 3.4 and 3.5.

Table 3.4. Winning history for Connecticut in 2010 season

Connecticut 2010	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Points
History results	Won	Won	Won	Won	Won	Won	
Single scoring	1	1	1	1	1	1	6
Double scoring	1	2	4	8	16	32	63

Table 3.5. Winning history for Connecticut in 2009 season

Connecticut 2009	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Points
History Results	Won	Won	Won	Won	Won	Won	
Single scoring	1	1	1	1	1	1	6
Double scoring	1	2	4	8	16	32	63

It is noted that Connecticut received 12 points and 126 points under the single and double scoring system, respectively.

3.2.2. Develop models for the first round using differences in ranks of seasonal averages with single scoring system variable

3.2.2.1. Ordinary least squares regression model

Differences in the ranks of the seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and single scoring system variable X2.

The response variable for the ordinary least squares regression model was point spread in the order of the team of interest minus the opposing team. A positive point spread indicates a win for the team of interest and a negative value indicates a loss for the team of interest. There were 192 teams playing 96 games in first rounds of the tournaments in 2011, 2012 and 2013. For the first half of the games of the first round in the three years, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the other half of the games of the first round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers).

No intercept was included when developing the models because the models should give the same results regardless of the ordering of the teams in the model. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages for all the variables previously given in Table

3.1 (Set A) were considered for entry in the model. Single scoring system variable X2 was also included for entry in the model.

3.2.2.1.1. Development of ordinary least squares regression model for the first round

The ordinary least squares regression model to estimate the point spread for each game in the first round based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-0.11158\text{*Diff in Scoring Margin}) + (0.02418\text{*Diff in Three-Point Field Goals Per Game}) + (2.40833\text{*X2 (SINGLE)})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.6. Table 3.7 gives the steps associated with the stepwise selection technique and Table 3.8 shows the associated R-square values as variables are added to the model. The model with the 3 significant variables explains an estimated 60% of the variation in point spread.

Table 3.6. Point spread model parameter estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Scoring_Margin	1	-0.11158	0.02405	-4.64	<.0001	1.30598
Three_Point_Goals	1	0.02418	0.01090	2.22	0.0289	1.02804
SINGLE	1	2.40833	0.35459	6.79	<.0001	1.33775

Table 3.7. Summary of stepwise selection for point spread model

Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	SINGLE		1	0.4946	0.4946	22.2472	92.96	<.0001
2	Scoring_Margin		2	0.0890	0.5835	3.7868	20.08	<.0001
3	Three_Point_Goals		3	0.0209	0.6045	0.9688	4.93	0.0289

Table 3.8. Summary of R-squares value

Root MSE	13.78907	R-Square	0.6045
Dependent Mean	-4.53125	Adj R-Sq	0.5917
Coeff Var	-304.31050		

3.2.2.2. Logistic regression model

The logistic regression model was also fit to the data with the dependent variable recorded as ‘1’ for win and ‘0’ for loss for the team of interest. Team of interest was the stronger team (higher seed numbers) in half of the games and the weaker team (lower seed numbers) in the other half of the games.

The intercept was excluded during the development of the logistic regression model because the ordering of the teams in the model should not matter. Stepwise selection was used with an α value of 0.1 for both entry and exit when determining the significant variables in developing the logistic regression model. The differences of the seasonal averages for both teams for all previously mentioned variables listed in Table 3.1 (Set A) were considered for entry in the model. Single scoring system variable X2 was also considered to enter the model.

3.2.2.2.1. Development of logistic regression model for the first round

A logistic regression model to estimate the probability of the team of interest winning the game for each game in the first round was developed and found to be:

$$\pi(\text{Diff\_FGP, Diff\_TPG, Diff\_BLK, Diff\_Single})= \frac{e^{-0.00908*\text{Diff\_FGP}+0.00574*\text{Diff\_TPG}-0.00457*\text{Diff\_BLK}+0.2466*\text{Diff\_Single}}}{1+e^{-0.00908*\text{Diff\_FGP}+0.00574*\text{Diff\_TPG}-0.00457*\text{Diff\_BLK}+0.2466*\text{Diff\_Single}}}$$

Where π (Diff_FGP, Diff_TPG, Diff_BLK, Diff_Single) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Field Goal Percentage, difference of seasonal averages in Three Point Goals, difference of seasonal averages in Blocked Shots and difference of seasonal averages in single scoring system variable in model.

Table 3.9 shows the steps for the stepwise selection technique and Table 3.10 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.11 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.1675 indicating that there was no evidence to reject using the logistic regression model.

Table 3.9. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	SINGLE		1	1	28.2358		<.0001
2	Three_Point_Goals		1	2	8.9312		0.0028
3	Field_Goal_PCT		1	3	7.4522		0.0063
4	Blocked_shots		1	4	4.1107		0.0426

Table 3.10. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Field_Goal_PCT	1	-0.00908	0.00349	6.7658	0.0093
Three_Point_Goals	1	0.00574	0.00231	6.1900	0.0128
Blocked_shots	1	-0.00457	0.00232	3.8638	0.0493
SINGLE	1	0.2466	0.0891	7.6649	0.0056

Table 3.11. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
11.6508	8	0.1675

3.2.3. Develop models for the first round using differences in ranks of seasonal averages with double scoring system variable

Difference in the ranks of seasonal averages for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A) were collected. The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and double scoring system variable X1.

The response variable for the ordinary least squares regression model was point spread in the order of the team of interest minus the opposing team. A positive point spread indicates a win for the team of interest and a negative value indicates a loss for the team of interest. There were 192 teams playing 96 games in first rounds of the tournaments in 2011, 2012 and 2013. For the first half games of the first round in the three years, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the other half of the games for the first round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers).

No intercept was included when developing the models because the models should give the same results regardless of the ordering of the teams in the model. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages for all the variables previously given in Table 3.1 (Set A) were considered for entry in the model. Double scoring system variable X1 was also considered to enter the model.

3.2.3.1. Ordinary least squares regression model

3.2.3.1.1. Development of ordinary least squares regression model for the first round

The ordinary least squares regression model to estimate the point spread for each game in the first round based on using differences between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-0.09799 * \text{Diff in Scoring Margin}) + (0.02447 * \text{Diff in Three-Point Field Goals Per Game}) + (-0.02182 * \text{Diff in Blocked Shots Per Game}) + (0.31844 * \text{X1 (DOUBLE)})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.12. Table 3.13 gives the steps associated with the stepwise selection technique and Table 3.14 shows the associated R-square values as variables are added to the model. The model with the 4 significant variables explains an estimated 56% of the variation in point spread.

Table 3.12. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Scoring_Margin	1	-0.09799	0.02761	-3.55	0.0006	1.54167
Three_Point_Goals	1	0.02447	0.01186	2.06	0.0419	1.09047
Blocked_shots	1	-0.02182	0.01165	-1.87	0.0643	1.34494
DOUBLE	1	0.31844	0.06432	4.95	<.0001	1.49309

Table 3.13. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	DOUBLE		1	0.4289	0.4289	23.5971	71.35	<.0001
2	Scoring_Margin		2	0.0834	0.5123	8.4261	16.07	0.0001
3	Three_Point_Goals		3	0.0339	0.5462	3.4426	6.95	0.0098
4	Blocked_shots		4	0.0167	0.5629	2.0106	3.51	0.0643

Table 3.14. Summary of R-squares value

Root MSE	14.57469	R-Square	0.5629
Dependent Mean	-4.53125	Adj R-Sq	0.5439
Coeff Var	-321.64826		

3.2.3.2. Logistic regression model

3.2.3.2.1. Development of logistic regression model for the first round

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the first round was developed and found to be:

$$\pi(\text{Diff\_FGP, Diff\_TPG, Diff\_BLK, Diff\_Double})= \frac{e^{-0.00899*\text{Diff\_FGP}+0.0056*\text{Diff\_TPG}-0.00449*\text{Diff\_BLK}+0.075*\text{Diff\_Double}}}{1+e^{-0.00899*\text{Diff\_FGP}+0.0056*\text{Diff\_TPG}-0.00449*\text{Diff\_BLK}+0.075*\text{Diff\_Double}}}$$

Where π (Diff_FGP, Diff_TPG, Diff_BLK, Diff_Double) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Field Goal

Percentage, difference of seasonal averages in Three Point Goals, difference of seasonal averages in Blocked shots and difference of seasonal averages in double scoring system variable in model.

Table 3.15 shows the steps for the stepwise selection technique and Table 3.16 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.17 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.2003 indicating that there was no evidence to reject using the logistic regression model.

Table 3.15. Summary of stepwise selection for logistic regression model

Step	**Effect**		**DF**	**Number**	**Score**	**Wald**	**Pr > ChiSq**
	Entered	**Removed**		**In**	**Chi-Square**	**Chi-Square**	
1	DOUBLE		1	1	22.0756		<.0001
2	Three_Point_Goals		1	2	8.5188		0.0035
3	Field_Goal_PCT		1	3	7.0184		0.0081
4	Blocked_shots		1	4	3.8729		0.0491

Table 3.16. Logistic regression model parameter estimates

Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Field_Goal_PCT	1	-0.00899	0.00351	6.5530	0.0105
Three_Point_Goals	1	0.00560	0.00233	5.7894	0.0161
Blocked_shots	1	-0.00449	0.00235	3.6478	0.0561
DOUBLE	1	0.0750	0.0314	5.6935	0.0170

Table 3.17. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	**DF**	**Pr > ChiSq**
11.0241	8	0.2003

3.2.4. Develop models for the second round using differences in ranks of seasonal averages with single scoring system variable

3.2.4.1. Ordinary least squares regression model

Rank differences based on seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included:

Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and single scoring system variable X2.

There were 96 teams playing 48 games in second rounds of the tournaments in 2011 to 2013. For the first half games of the second round, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages of the previously mentioned variables in Table 3.1 (Set A) were considered for entry in the model. Single scoring system variable was considered to enter the model. Single scoring system variable X2 was also considered to enter the model.

3.2.4.1.1. Development of ordinary least squares regression model for the second round

The ordinary least squares regression model to estimate the point spread for each game in the second round based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-0.108\text{*Diff in Scoring Margin}) + (-\ 0.03852\text{*Diff in Assist Turnover Ratio}) + (1.38319\text{*X2 (SINGLE)})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.18. Table 3.19 gives the steps associated with the stepwise selection technique and Table 3.20 shows associated R-square values as variables are added to the model. The model with the 3 significant variables explains an estimated 61% of the variation in point spread.

Table 3.18. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Scoring_Margin	1	-0.10800	0.03947	-2.74	0.0089	1.61525
Assist_Turnover_Ratio	1	-0.03852	0.01806	-2.13	0.0384	1.20444
SINGLE	1	1.38319	0.36197	3.82	0.0004	1.50182

Table 3.19. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	SINGLE		1	0.4698	0.4698	10.1554	41.64	<.0001
2	Scoring_Margin		2	0.1046	0.5744	1.0797	11.30	0.0016
3	Assist_Turnover_Ratio		3	0.0391	0.6134	-1.0583	4.55	0.0384

Table 3.20. Summary of R-squares value

Root MSE	10.54685	R-Square	0.6134
Dependent Mean	-1.06250	Adj R-Sq	0.5877
Coeff Var	-992.64498		

3.2.4.2. Logistic regression model

The logistic regression model was also fit for the data with responses recorded as‘1’ for win and ‘0’ for loss for the team of interest. This model estimates the probability of a win for the team of interest. No intercept was used during the development of the logistic regression model. Stepwise selection was used with an α value of 0.1 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences between the two teams of the seasonal averages of all previously mentioned variables in Table 3.1 (Set A) were considered for entry in the model. Single scoring system variable X2 was considered to enter the model.

3.2.4.2.1. Development of logistic regression model for the second round

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the second round was developed and found to be:

$$\pi(\text{Diff\_FGPD, Diff\_ATR, Diff\_TPFGD, Diff\_Single})= \frac{e^{0.0343*\text{Diff\_FGPD}-0.0409*\text{Diff\_ATR}-0.0262*\text{Diff\_TPFGD}+0.5967*\text{Diff\_Single}}}{1+e^{0.0343*\text{Diff\_FGPD}-0.0409*\text{Diff\_ATR}-0.0262*\text{Diff\_TPFGD}+0.5967*\text{Diff\_Single}}}$$

Where π (Diff_FGPD, Diff_ATR, Diff_TPFGD, Diff_Single) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Field Goal Percentage Defense, difference of seasonal averages in Assists Turnover Ratio, difference of seasonal averages in Three Point Field Goals Defense and difference of seasonal averages in single scoring system variable in model.

Table 3.21 shows the steps for the stepwise selection technique and Table 3.22 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.23 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.4236 indicating that there was no evidence to reject using the logistic regression model.

Table 3.21. Summary of stepwise selection for logistic regression model

Step	**Effect**		**DF**	**Number In**	**Score Chi-Square**	**Wald Chi-Square**	**Pr > ChiSq**
	Entered	**Removed**					
1	SINGLE		1	1	13.8737		0.0002
2	Assist_Turnover_Rati		1	2	6.9077		0.0086
3	Assists		1	3	3.9058		0.0481
4	Field_Goal_PCT_Dfens		1	4	5.9259		0.0149
5	Three_Pt_FG_Defense		1	5	13.5136		0.0002
6		Assists	1	4		2.0205	0.1552

Table 3.22. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Field_Goal_PCT_Dfens	1	0.0343	0.0128	7.2293	0.0072
Assist_Turnover_Rati	1	-0.0409	0.0184	4.9478	0.0261
Three_Pt_FG_Defense	1	-0.0262	0.00968	7.3448	0.0067
SINGLE	1	0.5967	0.1994	8.9554	0.0028

Table 3.23. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
8.1021	8	0.4236

3.2.5. Develop models for the second round using differences in ranks of seasonal averages with double scoring system variable

Rank differences of seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and double scoring system variable X1.

There were 96 teams playing 48 games in second rounds of the tournaments in 2011 to 2013. For the first half games of the second round, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages of the previously mentioned variables were considered for entry in the model. Double scoring system variable X1 was considered to enter the model.

3.2.5.1. Ordinary least squares regression model

3.2.5.1.1. Development of ordinary least squares regression model for the second round

The ordinary least squares regression model to estimate the point spread for each game in the second round based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-0.10627*\text{Diff in Scoring Margin}) + (0.03144*\text{Diff in Three-Point Field-Goal Percentage}) + (-0.03812*\text{Diff in Assist Turnover Ratio}) + (0.22394*\text{X1 (DOUBLE)})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.24. Table 3.25 gives the steps associated with the stepwise selection technique and Table 3.26 shows the associated R-square values as variables are added to the model. The model with the 4 significant variables explains an estimated 65% of the variation in point spread.

Table 3.24. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Scoring_Margin	1	-0.10627	0.03813	-2.79	0.0078	1.63111
Three_Point_Goal_PCT	1	0.03144	0.01608	1.96	0.0569	1.14973
Assist_Turnover_Ratio	1	-0.03812	0.01744	-2.19	0.0342	1.21532
DOUBLE	1	0.22394	0.05023	4.46	<.0001	1.66101

Table 3.25. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	DOUBLE		1	0.4923	0.4923	7.8769	45.58	<.0001
2	Scoring_Margin		2	0.0933	0.5856	-0.0225	10.35	0.0024
3	Assist_Turnover_Ratio		3	0.0347	0.6203	-1.7058	4.11	0.0485
4	Three_Point_Goal_PCT		4	0.0304	0.6507	-2.9282	3.82	0.0569

Table 3.26. Summary of R-squares value

Root MSE	10.13900	R-Square	0.6507
Dependent Mean	-1.06250	Adj R-Sq	0.6189
Coeff Var	-954.25902		

3.2.5.2. Logistic regression model

3.2.5.2.1. Development of logistic regression model for the second round

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the second round was developed and found to be:

$$\pi(\text{Diff\_FGPD, Diff\_ATR, Diff\_TPFGD, Diff\_Double})= \frac{e^{0.0259*\text{Diff\_FGPD}-0.03*\text{Diff\_ATR}-0.0202*\text{Diff\_TPFGD}+0.072*\text{Diff\_Double}}}{1+e^{0.0259*\text{Diff\_FGPD}-0.03*\text{Diff\_ATR}-0.0202*\text{Diff\_TPFGD}+0.072*\text{Diff\_Double}}}$$

Where π (Diff_FGPD, Diff_ATR, Diff_TPFGD, Diff_Double) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Field Goal Percentage Defense, difference of seasonal averages in Assists Turnover Ratio, difference of seasonal averages in Three Point Field Goals Defense and difference of seasonal averages in double scoring system variable in model.

Table 3.27 shows the steps for the stepwise selection technique and Table 3.28 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.29 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.0003 which is less than 0.5 indicating that there was evidence to reject using the logistic regression model.

Table 3.27. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	DOUBLE		1	1	11.5750		0.0007
2	Assist_Turnover_Rati		1	2	6.9345		0.0085
3	Three_Pt_FG_Defense		1	3	4.0577		0.0440
4	Field_Goal_PCT_Dfens		1	4	11.8057		0.0006

Table 3.28. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Field_Goal_PCT_Dfens	1	0.0259	0.00941	7.5523	0.0060
Assist_Turnover_Rati	1	-0.0300	0.0125	5.7261	0.0167
Three_Pt_FG_Defense	1	-0.0202	0.00705	8.2136	0.0042
DOUBLE	1	0.0720	0.0246	8.5571	0.0034

Table 3.29. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
29.1145	8	0.0003

3.2.6. Develop models for the third and higher rounds using differences in ranks of seasonal averages with single scoring system variable

3.2.6.1. Ordinary least squares regression model

Rank differences of seasonal averages for all the teams in the 2011, 2012 and 2013 tournaments were collected on the variables list in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and single scoring system variable X2.

There were 90 teams playing 45 games in second rounds of the tournaments in 2011 to 2013. For the first 20 games of the second round, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages of the previously mentioned variables were considered for entry in the model. Single scoring system variable X2 was also considered to enter the model.

3.2.6.1.1. Development of ordinary least squares regression model for the third and higher rounds

The ordinary least squares regression model to estimate the point spread for each game in the third and higher rounds based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-0.07693\text{*Diff in Scoring Defense}) + (-0.10473\text{*Diff in Assists Per Game})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.30. Table 3.31 gives the steps associated with the stepwise selection technique and Table 3.32 shows the associated R-square values as variables are added to the model. The model with the 2 significant variables explains an estimated 42% of the variation in point spread.

Table 3.30. Point spread model parameter estimates

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Scoring_Defense	-0.07693	0.02609	1843.66219	8.69	0.0051
Assists	-0.10473	0.02390	4074.00461	19.21	<.0001

Table 3.31. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Assists		1	0.2996	0.2996	21.9965	18.82	<.0001
2	Scoring_Defense		2	0.1178	0.4174	13.0669	8.69	0.0051

Table 3.32. Summary of R-squares value

Root MSE	14.56370	R-Square	0.4174
Dependent Mean	-3.86667	Adj R-Sq	0.3903
Coeff Var	-376.64740		

3.2.6.2. Logistic regression model

The logistic regression model was also fit for the data with responses recorded as'1' for win and '0' for loss for the team of interest. This model estimates the probability of a win for the team of interest. No intercept was used during the development of the logistic regression model. Stepwise selection was used with an α value of 0.1 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences between the two teams of the seasonal averages of all previously mentioned variables in Table 3.1 (Set A) were considered for entry in the model. Single scoring system variable X2 was also considered to enter the model.

3.2.6.2.1. Development of logistic regression model for the third and higher rounds

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the third and higher rounds was developed and found to be:

$$\pi(\text{Diff\_SO, Diff\_SD, Diff\_SM, Diff\_STL}) = \frac{e^{-0.0757*\text{Diff\_SO}-0.0229*\text{Diff\_SD}+0.0582*\text{Diff\_SM}-0.00585*\text{Diff\_STL}}}{1+e^{-0.0757*\text{Diff\_SO}-0.0229*\text{Diff\_SD}+0.0582*\text{Diff\_SM}-0.00585*\text{Diff\_STL}}}$$

Where π (Diff_SO, Diff_SD, Diff_SM, Diff_STL) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Scoring Offense, Difference of seasonal averages in Scoring Defense, difference of seasonal averages in Scoring Margin and difference of seasonal averages in Steals in model.

Table 3.33 shows the steps for the stepwise selection technique and Table 3.34 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.35 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.4358 indicating that there was no evidence to reject using the logistic regression model.

Table 3.33. Summary of stepwise selection for logistic regression model

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Scoring_Offense		1	1	9.3923		0.0022
2	Scoring_Defense		1	2	5.5685		0.0183
3	Scoring_Margin		1	3	3.6053		0.0576
4	Steals		1	4	3.0988		0.0783
5	Three_Point_Goal_PCT		1	5	2.8189		0.0932
6		Three_Point_Goal_PCT	1	4		2.5291	0.1118

Table 3.34. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Scoring_Offense	1	-0.0757	0.0319	5.6360	0.0176
Scoring_Defense	1	-0.0229	0.00948	5.8578	0.0155
Scoring_Margin	1	0.0582	0.0285	4.1742	0.0410
Steals	1	-0.00585	0.00346	2.8513	0.0913

Table 3.35. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
6.9342	7	0.4358

3.2.7. Develop models for the third and higher rounds using differences in ranks of seasonal averages with double scoring system variable

3.2.7.1. Ordinary least squares regression model

Differences in ranks of seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked

Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and double scoring system variable X1.

There were 90 teams playing 45 games in second rounds of the tournaments in 2011 to 2013. For the first 20 games of the second round, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages of the previously mentioned variables in Table 3.1 (Set A) were considered for entry in the model. Double scoring system variable X1 was also considered to enter the model.

3.2.7.1.1. Development of ordinary least squares regression model for the third and higher rounds

The ordinary least squares regression model to estimate the point spread for each game in the third and higher rounds based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-0.07693*\text{Diff in Scoring Defense}) + (-0.10473*\text{Diff in Assists Per Game})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.36. Table 3.37 gives the steps associated with the stepwise selection technique and Table 3.38 shows the associated R-square values as variables are added to the model. The model with the 2 significant variables explains an estimated 42% of the variation in point spread.

Table 3.36. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Scoring_Defense	1	-0.07693	0.02609	-2.95	0.0051	1.01018
Assists	1	-0.10473	0.02390	-4.38	<.0001	1.01018

Table 3.37. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Assists		1	0.2996	0.2996	23.1254	18.82	<.0001
2	Scoring_Defense		2	0.1178	0.4174	14.0060	8.69	0.0051

Table 3.38. Summary of R-squares value

Root MSE	14.56370	R-Square	0.4174
Dependent Mean	-3.86667	Adj R-Sq	0.3903
Coeff Var	-376.64740		

3.2.7.2. Logistic regression model

3.2.7.2.1. Development of logistic regression model for the third and higher rounds

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the third and higher rounds was developed and found to be:

$$\pi(\text{Diff\_SO, Diff\_SD, Diff\_SM, Diff\_STL})= \frac{e^{-0.0757*\text{Diff\_SO}-0.0229*\text{Diff\_SD}+0.0582*\text{Diff\_SM}-0.00585*\text{Diff\_STL}}}{1+e^{-0.0757*\text{Diff\_SO}-0.0229*\text{Diff\_SD}+0.0582*\text{Diff\_SM}-0.00585*\text{Diff\_STL}}}$$

Where π (Diff_SO, Diff_SD, Diff_SM, Diff_STL) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Scoring Offense, Difference of seasonal averages in Scoring Defense, difference of seasonal averages in Scoring Margin and difference of seasonal averages in Steals in model.

Table 3.39 shows the steps for the stepwise selection technique and Table 3.40 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.41 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.4358 indicating that there was no evidence to reject using the logistic regression model.

Table 3.39. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	Scoring_Offense		1	1	9.3923		0.0022
2	Scoring_Defense		1	2	5.5685		0.0183
3	Scoring_Margin		1	3	3.6053		0.0576
4	Steals		1	4	3.0988		0.0783
5	Three_Point_Goal_PCT		1	5	2.8189		0.0932
6		Three_Point_Goal_PCT	1	4		2.5291	0.1118

Table 3.40. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Scoring_Offense	1	-0.0757	0.0319	5.6360	0.0176
Scoring_Defense	1	-0.0229	0.00948	5.8578	0.0155
Scoring_Margin	1	0.0582	0.0285	4.1742	0.0410
Steals	1	-0.00585	0.00346	2.8513	0.0913

Table 3.41. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
6.9342	7	0.4358

3.2.8. Validating models

3.2.8.1. Validating first round using models developed with single scoring system variable

The 6 ordinary least squares regression models developed by using differences in ranks of seasonal averages data with either a single or double scoring system variable was used to validate the first round, second round and third round through final of 2014 season to check the validation accuracy of the models respectively. Logistic regression models were also used to do the validation but the results were not included in this paper since the accuracy is lower than ordinary least squares regression models. It is noted that the 2014 season was not used in the development of the models.

Table 3.42 gives the results as to how accurately the ordinary least squares regression model which developed by using differences in ranks of seasonal averages data with a single scoring system variable for first round of the NCAA 2014 women's basketball tournament.

Table 3.42. Accuracy of ordinary least squares regression model developed by differences in ranks of seasonal averages with a single scoring system variable when validating first round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	19	6	25
	Loss	5	2	7
	Total	24	8	32
Overall Accuracy				65.63%

3.2.8.2. Validating first round using models developed with double scoring system variable

Table 3.43 gives the results as to how accurately the ordinary least squares regression model which developed by using differences in ranks of seasonal averages data with a double scoring system variable for first round of the NCAA 2014 women's basketball tournament.

Table 3.43. Accuracy of ordinary least squares regression model developed by differences in ranks of seasonal averages with a double scoring system variable when validating first round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	21	4	25
	Loss	3	4	7
	Total	24	8	32
Overall Accuracy				78.13%

3.2.8.3. Validating second round using models developed with single scoring system variable

Table 3.44 gives the results as to how accurately the ordinary least squares regression model which developed by using differences in ranks of seasonal averages data with a single scoring system variable for second round of the NCAA 2014 women's basketball tournament.

Table 3.44. Accuracy of ordinary least squares regression model developed by differences in ranks of seasonal averages with a single scoring system variable when validating second round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	2	7
	Loss	1	7	8
	Total	6	9	16
Overall Accuracy				75%

3.2.8.4. Validating second round using models developed with double scoring system variable

Table 3.45 gives the results as to how accurately the ordinary least squares regression model which developed by using differences in ranks of seasonal averages data with a double scoring system variable for second round of the NCAA 2014 women's basketball tournament.

Table 3.45. Accuracy of ordinary least squares regression model developed by differences in ranks of seasonal averages with a double scoring system variable when validating second round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	3	8
	Loss	1	7	8
	Total	6	10	16
Overall Accuracy				75%

3.2.8.5. Validating third and higher rounds using models developed with single scoring system variable

Table 3.46 gives the results as to how accurately the ordinary least squares regression model which developed by using differences in ranks of seasonal averages data with a single scoring system variable for third and higher rounds of the NCAA 2014 women's basketball tournament.

Table 3.46. Accuracy of ordinary least squares regression model developed by differences in ranks of seasonal averages with a single scoring system variable when validating third and higher rounds of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	10	0	10
	Loss	1	4	5
	Total	11	4	15
Overall Accuracy				93.33%

3.2.8.6. Validating third and higher rounds using models developed with double scoring system variable

Table 3.47 gives the results as to how accurately the ordinary least squares regression model which developed by using differences in ranks of seasonal averages data with a double scoring system variable for third and higher rounds of the NCAA 2014 women's basketball tournament.

Table 3.47. Accuracy of ordinary least squares regression model developed by differences in ranks of seasonal averages with a double scoring system variable when validating third and higher rounds of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	10	0	10
	Loss	1	4	5
	Total	11	4	15
Overall Accuracy				93.33%

3.2.9. Bracketing the 2014 and 2015 tournament before tournament begins – Prediction

3.2.9.1. Using ordinary least squares regression models developed by differences in ranks of seasonal averages with a single scoring system variable

Results were predicted for every round before the tournament began. Significant differences in ranks of seasonal averages of variables were found for all teams playing in the first round and put into first round model. Significant differences of ranks of seasonal averages of variables were found for each team predicted to play each other in the second round were placed in second round model and winners of this round were predicted. Differences of ranks of seasonal averages of variables found to be significant of teams predicted to play each other in the third round were placed in the third round model and winning teams predicted for this round. This process continued until a winner is selected.

The predicted results were then compared against the actual results for every game in the 2014 and 2015 tournaments.

3.2.9.1.1. Examples for each round in 2014 tournament

An example for the first round, second round and then third or higher round will be given as to how the ordinary least squares regression model developed by using differences in ranks of seasonal averages with a single scoring system variable for a particular round in 2014 tournament was used.

3.2.9.1.1.1. Examples for seasonal averages models with single scoring system variable

3.2.9.1.1.1.1. Ordinary least squares regression model for first round

The ordinary least squares regression model for first round developed by using differences in ranks of seasonal averages with a single scoring system variable is:

$$\hat{Y} = (-0.11158*\text{Diff in Scoring Margin}) + (0.02418*\text{Diff in Three-Point Field Goals Per Game}) + (2.40833*\text{X2 (SINGLE)})$$

University of Connecticut played Prairie View in the first round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.48. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.48. University of Connecticut and Prairie View Statistics

Team	**Score**	**Scoring Margin***	**Three-Point Field Goals***	**Single scoring***
University of Connecticut	87	1	40	11
Prairie View	44	247	245	2
Difference	43	-246	-205	9

* Ranks based on seasonal averages

Using the model above, the game between University of Connecticut and Prairie View had a predicted point spread of:

$\hat{y}$ = (-0.11158*-246) + (0.02418*-205) + (2.40833*9) = 44.17

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for University of Connecticut, who won the game by a score of 87 to 44.

North Carolina State played BYU in the first round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.49. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.49. North Carolina State and BYU Statistics

Team	**Score**	**Scoring Margin***	**Three-Point Field Goals***	**Single scoring***
North Carolina State	57	44	46	0
BYU	72	61	101	1
Difference	-15	-17	-55	-1

* Ranks based on seasonal averages

Using the model above, the game between North Carolina State and BYU had a predicted point spread of:

$\hat{y}$ = (-0.11158*-17) + (0.02418*-55) + (2.40833*-1) = -1.84

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for North Carolina State, who lost the game by a score of 57 to 72.

DePaul played Oklahoma in the first round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.50. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.50. DePaul and Oklahoma Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Single scoring*
DePaul	104	21	15	3
Oklahoma	100	52	85	5
Difference	4	-31	-70	-2

* Ranks based on seasonal averages

Using the model above, the game between DePaul and Oklahoma had a predicted point spread of:

$\hat{y}$ = (-0.11158*-31) + (0.02418*-70) + (2.40833*-2) = -3.05

Since $\hat{y}$ <0 this game was coded as an incorrectly predicted loss for DePaul, who won the game by a score of 104 to 100.

Stanford played South Dakota in the first round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.51. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.51. Stanford and South Dakota Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Single scoring*
Stanford	81	9	79	8
South Dakota	62	123	127	1
Difference	19	-114	-48	7

* Ranks based on seasonal averages

Using the model above, the game between Stanford and South Dakota had a predicted point spread of:

$\hat{y}$ = (-0.11158*-114) + (0.02418*-48) + (2.40833*7) = 28.42

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Stanford, who won the game by a score of 81 to 62.

3.2.9.1.1.1.2. Ordinary least squares regression model for second round

The ordinary least squares regression model for second round developed by using differences in ranks of seasonal averages with a single scoring system variable is:

$$\hat{y} = (-0.108*\text{Diff in Scoring Margin}) + (-0.03852*\text{Diff in Assist Turnover Ratio}) + (1.38319*\text{X2 (SINGLE)})$$

Oklahoma State played Purdue in the second round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.52. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.52. Oklahoma State and Purdue Statistics

Team	Score	Scoring Margin*	Assist Turnover Ratio *	Single scoring*
Oklahoma State	73	32	77	2
Purdue	66	76	70	4
Difference	7	-44	7	-2

* Ranks based on seasonal averages

Using the model above, the game between Oklahoma State and Purdue had a predicted point spread of:

$$\hat{y} = (-0.108*-44) + (-0.03852*7) + (1.38319*-2) = 1.72$$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Oklahoma State, who won the game by a score of 73 to 66.

California played Baylor in the second round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.53. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.53. California and Baylor Statistics

Team	Score	Scoring Margin*	Assist Turnover Ratio *	Single scoring*
California	56	103	167	7
Baylor	75	3	3	9
Difference	-19	100	164	-2

* Ranks based on seasonal averages

Using the model above, the game between California and Baylor had a predicted point spread of:

$\hat{y}$ = (-0.108*100) + (-0.03852*164) + (1.38319*-2) = -19.88

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for California, who lost the game by a score of 56 to 75.

3.2.9.1.1.1.3. Ordinary least squares regression model for third and higher rounds

The ordinary least squares regression model for third and higher rounds developed by using differences in ranks of seasonal averages with a single scoring system variable is:

$\hat{y}$ = (-0.07693*Diff in Scoring Defense) + (-0.10473*Diff in Assists Per Game)

Notre Dame played Oklahoma State in the third round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.54.

Table 3.54. Notre Dame and Oklahoma State Statistics

Team	Score	Scoring Defense *	Assists *
Notre Dame	89	50	2
Oklahoma State	72	38	143
Difference	17	12	-141

* Ranks based on seasonal averages

Using the model above, the game between Notre Dame and Oklahoma State had a predicted point spread of:

$\hat{y}$ = (-0.07693*12) + (- 0.10473*-141) = 13.84

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Notre Dame, who won the game by a score of 89 to 72.

Tennessee played Maryland in the third round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.55.

Table 3.55. Tennessee and Maryland Statistics

Team	Score	Scoring Defense *	Assists *
Tennessee	62	79	47
Maryland	73	56	4
Difference	-11	23	43

* Ranks based on seasonal averages

Using the model above, the game between Tennessee and Maryland had a predicted point spread of:

$\hat{y} = (-0.07693*23) + (-0.10473*43) = -6.27$

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Tennessee, who lost the game by a score of 62 to 73.

3.2.9.2. Using models developed by differences in ranks of seasonal averages with a double scoring system variable

3.2.9.2.1. Examples for seasonal averages models with double scoring system variable

An example for the first round, second round and then third or higher round will be given as to how the ordinary least squares regression model developed by using differences in ranks of seasonal averages with a double scoring system variable for a particular round in 2014 tournament was used.

3.2.9.2.1.1. Ordinary least squares regression model for first round

The ordinary least squares regression model for first round developed by using differences in ranks of seasonal averages with a double scoring system variable is:

$$\hat{y} = (-0.09799*\text{Diff in Scoring Margin}) + (0.02447*\text{Diff in Three-Point Field Goals Per Game}) + (-0.02182*\text{Diff in Blocked Shots Per Game}) + (0.31844*\text{X1 (DOUBLE)})$$

University of Connecticut played Prairie View in the first round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.56. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.56. University of Connecticut and Prairie View Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Blocked shots	Double scoring*
University of Connecticut	87	1	40	1	94
Prairie View	44	247	245	70	2
Difference	43	-246	-205	-69	92

* Ranks based on seasonal averages

Using the model above, the game between University of Connecticut and Prairie View had a predicted point spread of:

$$\hat{y} = (-0.09799*-246) + (0.02447*-205) + (-0.02182*-69) + (0.31844*92) = 49.89$$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for University of Connecticut, who won the game by a score of 87 to 44.

North Carolina State played BYU in the first round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.57. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.57. North Carolina State and BYU Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Blocked shots	Double scoring*
North Carolina State	57	44	46	236	0
BYU	72	61	101	7	1
Difference	-15	-17	-55	229	-1

* Ranks based on seasonal averages

Using the model above, the game between North Carolina State and BYU had a predicted point spread of:

$\hat{y}$ = (-0.09799*-17) + (0.02447*-55) + (- 0.02182*229) + (0.31844*-1) = -4.99

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for North Carolina State, who lost the game by a score of 57 to 72.

DePaul played Oklahoma in the first round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.58. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.58. DePaul and Oklahoma Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Blocked shots	Double scoring*
DePaul	104	21	15	240	4
Oklahoma	100	52	85	124	10
Difference	4	-31	-70	116	-6

* Ranks based on seasonal averages

Using the model above, the game between DePaul and Oklahoma had a predicted point spread of:

$\hat{y}$ = (-0.09799*-31) + (0.02447*-70) + (- 0.02182*116) + (0.31844*-6) = -3.12

Since $\hat{y}$ <0 this game was coded as an incorrectly predicted loss for DePaul, who won the game by a score of 104 to 100.

Stanford played South Dakota in the first round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.59. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.59. Stanford and South Dakota Statistics

Team	**Score**	**Scoring Margin***	**Three-Point Field Goals***	**Blocked shots**	**Double scoring***
Stanford	81	9	79	93	38
South Dakota	62	123	127	223	1
Difference	19	-114	-48	-130	37

* Ranks based on seasonal averages

Using the model above, the game between Stanford and South Dakota had a predicted point spread of:

$$\hat{y} = (-0.09799*-114) + (0.02447*-48) +(- 0.02182*-130) + (0.31844*37) = 24.62$$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Stanford, who won the game by a score of 81 to 62.

3.2.9.2.1.2. Ordinary least squares regression model for second round

The ordinary least squares regression model for second round developed by using differences in ranks of seasonal averages with a double scoring system variable is:

$$\hat{y} = (-0.10627*\text{Diff in Scoring Margin}) + (0.03144*\text{Diff in Three-Point Field-Goal Percentage}) + (- 0.03812*\text{Diff in Assist Turnover Ratio}) + (0.22394*\text{X1 (DOUBLE)})$$

Oklahoma State played Purdue in the second round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.60. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.60. Oklahoma State and Purdue Statistics

Team	Score	Scoring Margin*	Three-Point Field-Goal Percentage	Assist Turnover Ratio *	Double scoring*
Oklahoma State	73	32	156	77	3
Purdue	66	76	6	70	6
Difference	7	-44	150	7	-3

* Ranks based on seasonal averages

Using the model above, the game between Oklahoma State and Purdue had a predicted point spread of:

$\hat{y}$ = (-0.10627*-44) + (0.03144*150) + (- 0.03812*7) + (0.22394*-3) = 8.45

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Oklahoma State, who won the game by a score of 73 to 66.

California played Baylor in the second round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.61. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.61. California and Baylor Statistics

Team	Score	Scoring Margin*	Three-Point Field-Goal Percentage	Assist Turnover Ratio *	Double scoring*
California	56	103	262	167	34
Baylor	75	3	88	3	70
Difference	-19	100	174	164	-36

* Ranks based on seasonal averages

Using the model above, the game between California and Baylor had a predicted point spread of:

$\hat{y}$ = (-0.10627*100) + (0.03144*174) + (- 0.03812*164) + (0.22394*-36) = -19.47

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for California, who lost the game by a score of 56 to 75.

3.2.9.2.1.3. Ordinary least squares regression model for third and higher rounds

The ordinary least squares regression model for third and higher rounds developed by using differences in ranks of seasonal averages with a double scoring system variable is:

$\hat{y}$ = (-0.07693*Diff in Scoring Defense) + (-0.10473*Diff in Assists Per Game)

Notre Dame played Oklahoma State in the third round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.62.

Table 3.62. Notre Dame and Oklahoma State Statistics

Team	Score	Scoring Defense *	Assists *
Notre Dame	89	50	2
Oklahoma State	72	38	143
Difference	17	12	-141

* Ranks based on seasonal averages

Using the model above, the game between Notre Dame and Oklahoma State had a predicted point spread of:

$\hat{y}$ = (-0.07693*12) + (- 0.10473*-141) = -18.85

Since $\hat{y}$ > 0 this game was coded as a correctly predicted win for Notre Dame, who won the game by a score of 89 to 72.

Tennessee played Maryland in the third round of the 2014 Tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 3.63.

Table 3.63. Tennessee and Maryland Statistics

Team	Score	Scoring Defense *	Assists *
Tennessee	62	79	47
Maryland	73	56	4
Difference	-11	23	43

* Ranks based on seasonal averages

Using the model above, the game between Tennessee and Maryland had a predicted point spread of:

$\hat{y}$ = (-0.07693*23) + (- 0.10473*43) = -6.27

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for Tennessee, who lost the game by a score of 62 to 73.

3.2.10. Results for prediction

3.2.10.1. Results for prediction when using models developed using differences in ranks of seasonal averages with a single scoring system variable

In 2014, a continuous process was used in verifying the models instead of doing round by round predictions as in previous chapter. In other words, a complete bracket was filled out in 2014 before any game was played.

The ordinary least squares regression model for the first round developed by using differences in ranks of seasonal averages and a single scoring system variable was used to predict the teams who go to next round. Once the teams in the second round were predicted, the second-round models were used to predict the winners of the second round. This process was continued for the third and higher rounds until the predicted final winner of the game was determined.

A summary of the number of correct and incorrect predictions for each round of the 2014 tournament is given in Table 3.64.

Table 3.64. Prediction results of each round for 2014: (Ordinary least squares regression model developed by differences in ranks of seasonal averages with a single scoring system variable)

	Correct	Incorrect	Total games
First round	21	11	32
Second round	10	6	16
Third round	5	3	8
Fourth round	4	0	4
Fifth round	2	0	2
Final round	1	0	1
	Overall Accuracy		68.25%

A similar process was conducted to verifying the models for 2015 season. Namely, a complete bracket was filled out before 2015 tournament started.

A summary of the number of correct and incorrect predictions for each round of the 2015 tournament is given in Table 3.65.

Table 3.65. Prediction results of each round for 2015: (Ordinary least squares regression model developed by differences in ranks of seasonal averages with a single scoring system variable)

	Correct	Incorrect	Total games
First round	24	8	32
Second round	13	3	16
Third round	6	2	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	1	0	1
	Overall Accuracy		74.6%

3.2.10.2. Results for prediction when using models developed using differences in ranks of seasonal averages with a double scoring system variable

A similar process was conducted as in the previous section using the models developed by using differences in ranks of seasonal averages and a double scoring system variable to predict the results of the 2014 tournament. A complete bracket was filled out in 2014 before any game was played.

The ordinary least squares regression model for the first round developed by using differences in ranks of seasonal averages with a double scoring system variable was used to predict the teams who go to next round. Once the teams in the second round were predicted, the second-round models were used to predict the winners of the second round. This process was continued for the third and higher rounds until the predicted final winner of the game was determined.

A summary of the number of correct and incorrect predictions for each round of the 2014 tournament is given in Table 3.66.

Table 3.66. Prediction results of each round for 2014: (Ordinary least squares regression model developed by differences in ranks of seasonal averages with a double scoring system variable)

	Correct	Incorrect	Total games
First round	25	7	32
Second round	10	6	16
Third round	6	2	8
Fourth round	4	0	4
Fifth round	2	0	2
Final round	1	0	1
	Overall Accuracy		76.19%

A similar process was conducted to verifying the models for 2015 season. Namely, a complete bracket was filled out before 2015 tournament started.

A summary of the number of correct and incorrect predictions of the ordinary least squares regression model for each round of the 2015 tournament is given in Table 3.67.

Table 3.67. Prediction results of each round for 2015: (Ordinary least squares regression model developed by differences in ranks of seasonal averages with a double scoring system variable)

	Correct	Incorrect	Total games
First round	25	7	32
Second round	11	5	16
Third round	6	2	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	1	0	1
	Overall Accuracy		73.02%

It is noted that ordinary least squares regression models developed by using differences in ranks of seasonal averages with a double scoring system variable gave better results than the ordinary least squares regression models developed by using differences in ranks of seasonal averages with a single scoring system variable.

3.3 Develop models using differences of seasonal averages

3.3.1. Develop models for the first round using differences of seasonal averages with single scoring system variable

Seasonal averages from each of the team playing in the tournament were collected from the official NCAA basketball statistics database (NCAA [3]). Data was collected for three years of the NCAA women's basketball tournament and it was collected before the tournament started. For example, the first game of NCAA 2011 women's basketball tournament began on March 19, 2011, all the data was collected through games March 16, 2011. This included 2011, 2012 and 2013 tournaments. Seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and single scoring system variable X2.

The response variable for the ordinary least squares regression model was point spread in the order of the team of interest minus the opposing team. A positive point spread indicates a win for the team of interest and a negative value indicates a loss for the team of interest. There were 192 teams playing 96 games in first rounds of the tournaments in 2011, 2012 and 2013. For the first 48 games of the first round in the three years, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the other 48 games of the first round of the three years, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers).

No intercept was included when developing the models because the models should give the same results regardless of the ordering of the teams in the model. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages for all the variables previously given in Table 3.1 (Set A) were considered for entry in the model. Single scoring system variable X2 was also considered to enter the model.

3.3.1.1. Ordinary least squares regression model

3.3.1.1.1 Development of ordinary least squares regression model for the first round

The ordinary least squares regression model to estimate the point spread for each game in the first round based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (0.78646 \text{*Diff in Scoring Margin}) + (-\ 1.75276 \text{*Diff in Three - Point Field Goals Per Game}) + (2.0739 \text{*X2 (SINGLE)})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.68. Table 3.69 gives the steps associated with the stepwise selection technique and Table 3.70 shows the associated R-square values as variables are added to the model. The model with the 3 significant variables explains an estimated 62% of the variation in point spread.

Table 3.68. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Scoring_Margin	1	0.78646	0.15284	5.15	<.0001	1.48851
Three_Point_Goals	1	-1.75276	0.66810	-2.62	0.0102	1.05068
SINGLE	1	2.07390	0.37320	5.56	<.0001	1.54867

Table 3.69. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	SINGLE		1	0.4946	0.4946	19.5703	92.96	<.0001
2	Scoring_Margin		2	0.0990	0.5935	-0.6636	22.88	<.0001
3	Three_Point_Goals		3	0.0280	0.6215	-4.9574	6.88	0.0102

Table 3.70. Summary of R-squares value

Root MSE	13.48857	R-Square	0.6215
Dependent Mean	-4.53125	Adj R-Sq	0.6093
Coeff Var	-297.67880		

3.3.1.2. Logistic regression model

3.3.1.2.1. Development of logistic regression model for the first round

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the first round was developed and found to be:

$$\pi(\text{Diff\_TPG, Diff\_FGP, Diff\_Single})=\frac{e^{-0.4522*\text{Diff\_TPG}+0.2614*\text{Diff\_FGP}+0.2325*\text{Diff\_Single}}}{1+e^{-0.4522*\text{Diff\_TPG}+0.2614*\text{Diff\_FGP}+0.2325*\text{Diff\_Single}}}$$

Where π (Diff_TPG, Diff_FGP, Diff_Single) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Three Point Goals, difference of seasonal averages in Field Goal Percentage and difference of seasonal averages in Single Scoring system variable in model.

Table 3.71 shows the steps for the stepwise selection technique and Table 3.72 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.73 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.4344 indicating that there was no evidence to reject using the logistic regression model.

Table 3.71. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	SINGLE		1	1	28.2358		<.0001
2	Three_Point_Goals		1	2	9.7432		0.0018
3	Field_Goal_PCT		1	3	9.0850		0.0026

Table 3.72. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Field_Goal_PCT	1	0.2614	0.0918	8.1154	0.0044
Three_Point_Goals	1	-0.4522	0.1493	9.1798	0.0024
SINGLE	1	0.2325	0.0905	6.6025	0.0102

Table 3.73. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
7.9904	8	0.4344

3.3.2. Develop models for the first round using differences in seasonal averages with double scoring system variable

Seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included: Scoring

Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and double scoring system variable X1.

The response variable for the ordinary least squares regression model was point spread in the order of the team of interest minus the opposing team. A positive point spread indicates a win for the team of interest and a negative value indicates a loss for the team of interest. There were 192 teams playing 96 games in first rounds of the tournaments in 2011, 2012 and 2013. For the first 48 games of the first round in the three years, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the other 48 games of the first round of the three years, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers).

No intercept was included when developing the models because the models should give the same results regardless of the ordering of the teams in the model. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages for all the variables previously given in Table 3.1 (Set A) were considered for entry in the model. Double scoring system variable X1 was also considered to enter the model.

3.3.2.1. Ordinary least squares regression model

3.3.2.1.1. Development of ordinary least squares regression model for the first round

The ordinary least squares regression model to estimate the point spread for each game in the first round based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (0.73021*\text{Diff in Scoring Margin}) + (-\ 1.8507*\text{Diff in Three-Point Field Goals Per Game}) + (1.62689*\text{Diff in Blocked shots}) + (0.2645*\text{X1 (DOUBLE)})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.74. Table 3.75 gives the steps associated with the stepwise selection technique and Table 3.76 shows the associated R-square values as variables are added to the model. The model with the 4 significant variables explains an estimated 59% of the variation in point spread.

Table 3.74. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Scoring_Margin	1	0.73021	0.17768	4.11	<.0001	1.83273
Three_Point_Goals	1	-1.85070	0.71798	-2.58	0.0115	1.10553
Blocked_shots	1	1.62689	0.97048	1.68	0.0971	1.43301
DOUBLE	1	0.26450	0.06636	3.99	0.0001	1.69081

Table 3.75. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Scoring_Margin		1	0.4335	0.4335	25.8626	72.70	<.0001
2	DOUBLE		2	0.0978	0.5313	7.1647	19.62	<.0001
3	Three_Point_Goals		3	0.0452	0.5765	-0.3933	9.92	0.0022
4	Blocked_shots		4	0.0126	0.5891	-1.0493	2.81	0.0971

Table 3.76. Summary of R-squares value

Root MSE	14.13146	R-Square	0.5891
Dependent Mean	-4.53125	Adj R-Sq	0.5712
Coeff Var	-311.86662		

3.3.2.2. Logistic regression model

3.3.2.2.1. Development of logistic regression model for the first round

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the first round was developed and found to be:

$$\pi(\text{Diff\_TPG, Diff\_FGP, Diff\_Double}) = \frac{e^{-0.4381*\text{Diff\_TPG}+0.2523*\text{Diff\_FGP}+0.0749*\text{Diff\_Double}}}{1+e^{-0.4381*\text{Diff\_TPG}+0.2523*\text{Diff\_FGP}+0.0749*\text{Diff\_Double}}}$$

Where π (Diff_TPG, Diff_FGP, Diff_Double) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Three Point Goals, difference of seasonal averages in Field Goal Percentage and difference of seasonal averages in double scoring system variable in model.

Table 3.77 shows the steps for the stepwise selection technique and Table 3.78 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.79 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.0511 indicating that there was no evidence to reject using the logistic regression model.

Table 3.77. Summary of stepwise selection for logistic regression model

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Field_Goal_PCT		1	1	24.8076		<.0001
2	Three_Point_Goals		1	2	15.4098		<.0001
3	DOUBLE		1	3	6.3523		0.0117

Table 3.78. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Field_Goal_PCT	1	0.2523	0.0924	7.4605	0.0063
Three_Point_Goals	1	-0.4381	0.1505	8.4696	0.0036
DOUBLE	1	0.0749	0.0323	5.3831	0.0203

Table 3.79. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
15.4400	8	0.0511

3.3.3. Develop models for the second round using differences in seasonal averages with single scoring system variable

Seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and single scoring system variable X2.

There were 96 teams playing 48 games in second rounds of the tournaments in 2011 to 2013. For the first 24 games of the second round, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the other 24 games of the second round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages of the previously mentioned variables in Table 3.1 (Set A) were considered for entry in the model. Single scoring system variable X2 was considered to enter the model.

3.3.3.1. Ordinary least squares regression model

3.3.3.1.1. Development of ordinary least squares regression model for the second round

The ordinary least squares regression model to estimate the point spread for each game in the second round based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (1.3091 * \text{Diff in Scoring Margin}) + (-\ 0.33382 * \text{Diff in Won-Lost Percentage}) + (0.91812 * \text{X2 (SINGLE)})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.80. Table 3.81 gives the steps associated with the stepwise selection technique and Table 3.82 shows the associated R-square values as variables are added to the model. The model with the 3 significant variables explains an estimated 68% of the variation in point spread.

Table 3.80. Point spread model parameter estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Scoring_Margin	1	1.30910	0.28125	4.65	<.0001	5.25999
Won_lost_PCT	1	-0.33382	0.15593	-2.14	0.0377	4.19159
SINGLE	1	0.91812	0.36244	2.53	0.0149	1.81583

Table 3.81. Summary of stepwise selection for point spread model

Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Scoring_Margin		1	0.5941	0.5941	6.6798	68.79	<.0001
2	SINGLE		2	0.0527	0.6468	1.8388	6.87	0.0119
3	Won_lost_PCT		3	0.0326	0.6795	-0.3984	4.58	0.0377

Table 3.82. Summary of R-squares value

Root MSE	9.60394	R-Square	0.6795
Dependent Mean	-1.06250	Adj R-Sq	0.6581
Coeff Var	-903.89998		

3.3.3.2. Logistic regression model

3.3.3.2.1. Development of logistic regression model for the second round

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the second round was developed and found to be:

$$\pi(\text{Diff\_SO, Diff\_TP}) = \frac{e^{0.1752*\text{Diff\_SO}-0.2899*\text{Diff\_TP}}}{1+e^{0.1752*\text{Diff\_SO}-0.2899*\text{Diff\_TP}}}$$

Where π (Diff_SO, Diff_TP) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Scoring Offense and difference of seasonal averages in Turnover in model.

Table 3.83 shows the steps for the stepwise selection technique and Table 3.84 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.85 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.8415 indicating that there was no evidence to reject using the logistic regression model.

Table 3.83. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	Scoring_Offense		1	1	15.6689		<.0001
2	Diff in Turnovers Pe		1	2	4.3404		0.0372
3	SINGLE		1	3	2.7360		0.0981
4		SINGLE	1	2		2.4507	0.1175

Table 3.84. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Scoring_Offense	1	0.1752	0.0555	9.9648	0.0016
Diff in Turnovers Pe	1	-0.2899	0.1506	3.7066	0.0542

Table 3.85. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
4.1692	8	0.8415

3.3.4. Develop models for the second round using differences in seasonal averages with double scoring system variable

Seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and double scoring system variable X1.

There were 96 teams playing 48 games in second rounds of the tournaments in 2011 to 2013. For the first 24 games of the second round, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the other 24 games of the second round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages of the previously mentioned variables were considered for entry in the model. Double scoring system variable X1 was considered to enter the model.

3.3.4.1. Ordinary least squares regression model

3.3.4.1.1. Development of ordinary least squares regression model for the second round

The ordinary least squares regression model to estimate the point spread for each game in the second round based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (1.27219*\text{Diff in Scoring Margin}) + (-\ 0.3248*\text{Diff in Won-Lost Percentage}) + (0.13556*\text{X1 (DOUBLE)})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.86. Table 3.87 gives the steps associated with the stepwise selection technique and Table 3.88 shows the associated R-square values as variables are added to the model. The model with the 3 significant variables explains an estimated 69% of the variation in point spread.

Table 3.86. Point spread model parameter estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Scoring_Margin	1	1.27219	0.28124	4.52	<.0001	5.36207
Won_lost_PCT	1	-0.32480	0.15466	-2.10	0.0414	4.20411
DOUBLE	1	0.13556	0.04977	2.72	0.0091	1.85231

Table 3.87. Summary of stepwise selection for point spread model

Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Scoring_Margin		1	0.5941	0.5941	6.9950	68.79	<.0001
2	DOUBLE		2	0.0607	0.6548	1.0728	8.09	0.0066
3	Won_lost_PCT		3	0.0308	0.6856	-0.9504	4.41	0.0414

Table 3.88. Summary of R-squares value

Root MSE	9.51159	R-Square	0.6856
Dependent Mean	-1.06250	Adj R-Sq	0.6646
Coeff Var	-895.20832		

3.3.4.2. Logistic regression model

3.3.4.2.1. Development of logistic regression model for the second round

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the second round was developed and found to be:

$$\pi(\text{Diff\_SO, Diff\_TP}) = \frac{e^{0.1752*\text{Diff\_SO}-0.2899*\text{Diff\_TP}}}{1+e^{0.1752*\text{Diff\_SO}-0.2899*\text{Diff\_TP}}}$$

Where π (Diff_SO, Diff_TP) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Scoring Offense and difference of seasonal averages in Turnover in model.

Table 3.89 shows the steps for the stepwise selection technique and Table 3.90 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.91 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.8415 indicating that there was no evidence to reject using the logistic regression model.

Table 3.89. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	Scoring_Offense		1	1	15.6689		<.0001
2	Diff in Turnovers Pe		1	2	4.3404		0.0372

Table 3.90. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Scoring_Offense	1	0.1752	0.0555	9.9648	0.0016
Diff in Turnovers Pe	1	-0.2899	0.1506	3.7066	0.0542

Table 3.91. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
4.1692	8	0.8415

3.3.5. Develop models for the third and higher rounds using differences in seasonal averages with single scoring system variable

Seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and single scoring system variable X2.

There were 90 teams playing 45 games in second rounds of the tournaments in 2011 to 2013. For the first 20 games of the second round, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages of the previously mentioned variables in Table 3.1 (Set A) were considered for entry in the model. Single scoring system variable X2 was also considered to enter the model.

3.3.5.1. Ordinary least squares regression model

3.3.5.1.1. Development of ordinary least squares regression model for the third and higher rounds

The ordinary least squares regression model to estimate the point spread for each game in the third and higher rounds based on using difference between seasonal averages of the significant variables was developed and found to be:

$\hat{Y}$ = (1.9762*Diff in Scoring Margin) + (- 0.71222*Diff in Won-Lost Percentage)

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.92. Table 3.93 gives the steps associated with the stepwise selection technique and Table 3.94 shows the associated R-square values as variables are added to the model. The model with the 2 significant variables explains an estimated 52% of the variation in point spread.

Table 3.92. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Scoring_Margin	1	1.97620	0.31787	6.22	<.0001	2.89771
Won_lost_PCT	1	-0.71222	0.21693	-3.28	0.0020	2.89771

Table 3.93. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Scoring_Margin		1	0.4058	0.4058	2.9188	30.05	<.0001
2	Won_lost_PCT		2	0.1191	0.5249	-4.2849	10.78	0.0020

Table 3.94. Summary of R-squares value

Root MSE	13.15148	R-Square	0.5249
Dependent Mean	-3.86667	Adj R-Sq	0.5028
Coeff Var	-340.12437		

3.3.5.2. Logistic regression model

3.3.5.2.1. Development of logistic regression model for the third and higher rounds

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the third and higher rounds was developed and found to be:

$$\pi(\text{Diff\_SM, Diff\_PF}) = \frac{e^{0.225*\text{Diff\_SM}+0.3563*\text{Diff\_PF}}}{1+e^{0.225*\text{Diff\_SM}+0.3563*\text{Diff\_PF}}}$$

Where π (Diff_SM, Diff_PF) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Scoring Margin and difference of seasonal averages in Personal Fouls in model.

Table 3.95 shows the steps for the stepwise selection technique and Table 3.96 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.97 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.2853 indicating that there was no evidence to reject using the logistic regression model.

Table 3.95. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	Scoring_Margin		1	1	14.7294		0.0001
2	Personal_Fouls		1	2	3.9261		0.0475

Table 3.96. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Scoring_Margin	1	0.2250	0.0695	10.4707	0.0012
Personal_Fouls	1	0.3563	0.1910	3.4792	0.0621

Table 3.97. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
8.5670	7	0.2853

3.3.6. Develop models for the third and higher rounds using differences in seasonal averages with double scoring system variable

Seasonal averages were collected for all the teams in the 2011, 2012 and 2013 tournaments on the variables listed in Table 3.1 (Set A). The variables included: Scoring Offense, Scoring Defense, Scoring Margin, Field-Goal Percentage, Field-Goal Percentage Defense, Free-Throw Percentage, Rebound Margin, Three-Point Field Goals Per Game, Three-Point Field Goal Percentage, Won-Lost Percentage, Assists Per Game, Blocked Shots Per Game, Steals Per Game, Turnovers Per game, Personal Fouls Per Game, Assist Turnover Ratio, Turnover Margin, Three Pt FG Defense and double scoring system variable X1.

There were 90 teams playing 45 games in second rounds of the tournaments in 2011 to 2013. For the first 20 games of the second round, the point spread was obtained by using the scores of weaker teams (higher seed numbers) minus the scores of stronger teams (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the scores of stronger teams (lower seed numbers) minus the scores of weaker teams (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences between the two teams of the seasonal averages of the previously mentioned variables in Table 3.1 (Set A) were considered for entry in the model. Double scoring system variable X1 was also considered to enter the model.

3.3.6.1. Ordinary least squares regression model

3.3.6.1.1. Development of ordinary least squares regression model for the third and higher rounds

The ordinary least squares regression model to estimate the point spread for each game in the third and higher rounds based on using difference between seasonal averages of the significant variables was developed and found to be:

$\hat{Y}$ = (1.9762*Diff in Scoring Margin) + (- 0.71222*Diff in Won-Lost Percentage)

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.98. Table 3.99 gives the steps associated with the stepwise selection technique and Table 3.100 shows the associated R-square values as variables are added to the model. The model with the 2 significant variables explains an estimated 52% of the variation in point spread.

Table 3.98. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Scoring_Margin	1	1.97620	0.31787	6.22	<.0001	2.89771
Won_lost_PCT	1	-0.71222	0.21693	-3.28	0.0020	2.89771

Table 3.99. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Scoring_Margin		1	0.4058	0.4058	2.8156	30.05	<.0001
2	Won_lost_PCT		2	0.1191	0.5249	-4.3674	10.78	0.0020

Table 3.100. Summary of R-squares value

Root MSE	13.15148	R-Square	0.5249
Dependent Mean	-3.86667	Adj R-Sq	0.5028
Coeff Var	-340.12437		

3.3.6.2. Logistic regression model

3.3.6.2.1. Development of logistic regression model for the third and higher rounds

A logistic regression model to help estimate the probability of the team of interest winning the game for each game in the third and higher rounds was developed and found to be:

$$\pi(\text{Diff\_SM, Diff\_PF}) = \frac{e^{0.225*\text{Diff\_SM}+0.3563*\text{Diff\_PF}}}{1+e^{0.225*\text{Diff\_SM}+0.3563*\text{Diff\_PF}}}$$

Where π (Diff_SM, Diff_PF) is the estimated probability that the team of interest will win the game with difference of seasonal averages in Scoring Margin and difference of seasonal averages in Personal Fouls in model.

Table 3.101 shows the steps for the stepwise selection technique and Table 3.102 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.103 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.2853 indicating that there was no evidence to reject using the logistic regression model.

Table 3.101. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	Scoring_Margin		1	1	14.7294		0.0001
2	Personal_Fouls		1	2	3.9261		0.0475

Table 3.102. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Scoring_Margin	1	0.2250	0.0695	10.4707	0.0012
Personal_Fouls	1	0.3563	0.1910	3.4792	0.0621

Table 3.103. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
8.5670	7	0.2853

3.3.7. Validating models

3.3.7.1. Validating first round using models developed with single scoring system variable

The 6 ordinary least squares regression models developed by using seasonal averages data with either a single or double scoring system variable were validated using the first round, second round and third round through final of the 2014 tournament. It is noted that the 2014 season was not used in the development of the models.

When develop the logistic regression model by using differences in ranks of seasonal averages with a double scoring system variable, the p-value for Hosmer and Lemeshow Test [8] was 0.0003 indicating that there was evidence to reject using the logistic regression model.

When develop the logistic regression model by using differences of seasonal averages with a double scoring system variable, the p-value for Hosmer and Lemeshow Test [8] was 0.0511 indicating that there was evidence to reject using the logistic regression model.

Because the p-value for the Hosmer and Lemeshow Test was 0.0003 and 0.0511 when developing the models with the double and single scoring system variables respectively, the logistic regression models were not included in this research.

Table 3.104 gives the results as to how accurately the ordinary least squares regression model which developed by using seasonal averages data with a single scoring system variable for first round of the NCAA 2014 women's basketball tournament.

Table 3.104. Accuracy of ordinary least squares regression model developed by using seasonal averages with a single scoring system variable when validating first round of 2014

Point spread		Predicted		
		Win	**Loss**	**Total**
Actual	**Win**	19	6	25
	Loss	5	2	7
	Total	24	8	32
Overall Accuracy				65.63%

3.3.7.2. Validating first round using models developed with double scoring system variable

Table 3.105 gives the results as to how accurately the ordinary least squares regression model which developed by using seasonal averages data with a double scoring system variable for first round of the NCAA 2014 women's basketball tournament.

Table 3.105. Accuracy of ordinary least squares regression model developed by using seasonal averages with a double scoring system variable when validating first round of 2014

Point spread		Predicted		
		Win	**Loss**	**Total**
Actual	**Win**	20	5	25
	Loss	3	4	7
	Total	23	9	32
Overall Accuracy				75%

3.3.7.3. Validating second round using models developed with single scoring system variable

Table 3.106 gives the results as to how accurately the ordinary least squares regression model which developed by using seasonal averages data with a single scoring system variable for second round of the NCAA 2014 women's basketball tournament.

Table 3.106. Accuracy of ordinary least squares regression model developed by using seasonal averages with a single scoring system variable when validating second round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	3	8
	Loss	1	7	8
	Total	6	10	16
Overall Accuracy				75%

3.3.7.4. Validating second round using models developed with double scoring system variable

Table 3.107 gives the results as to how accurately the ordinary least squares regression model which developed by using seasonal averages data with a double scoring system variable for second round of the NCAA 2014 women's basketball tournament.

Table 3.107. Accuracy of ordinary least squares regression model developed by using seasonal averages with a double scoring system variable when validating second round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	3	8
	Loss	1	7	8
	Total	6	10	16
Overall Accuracy				75%

It is noted that both models developed by using either the double or single scoring system variable have the same accuracy in this case.

3.3.7.5. Validating third and higher rounds using models developed with single scoring system variable

Table 3.108 gives the results as to how accurately the ordinary least squares regression model which developed by using seasonal averages data with a single scoring system variable for third and higher rounds of the NCAA 2014 women's basketball tournament.

Table 3.108. Accuracy of ordinary least squares regression model developed by using seasonal averages with a single scoring system variable when validating third and higher rounds of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	8	2	10
	Loss	1	4	5
	Total	9	6	15
Overall Accuracy				80%

3.3.7.6. Validating third and higher rounds using models developed with double scoring system variable

Table 3.109 gives the results as to how accurately the ordinary least squares regression model which developed by using seasonal averages data with a double scoring system variable for third and higher rounds of the NCAA 2014 women's basketball tournament.

Table 3.109. Accuracy of ordinary least squares regression model developed by using seasonal averages with a double scoring system variable when validating third and higher rounds of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	8	2	10
	Loss	1	4	5
	Total	9	6	15
Overall Accuracy				80%

It is found that when seasonal averages were used in development of the models, the validation accuracy was the same for both models with single and double scoring system variables in this case.

3.3.8. Bracketing the 2015 tournament before tournament begins - Prediction (models developed by using seasonal averages with a single scoring system variable)

Results were predicted for every round before the tournament began. Differences of seasonal averages of variables to be significant were collected for all teams playing in the first round and put into the first round model. Significant differences of seasonal averages for each predicted to play each other in the second round were placed in second round model and winners of this round were predicted. Differences of seasonal averages of variables found to be significant of teams predicted to play each other in the third round were placed in the third round model and winning teams predicted for this round. This process continued until a winner is selected.

The predicted results were then compared against the actual results for each round of the game for 2014 and 2015.

3.3.8.1. Examples for seasonal averages models with single scoring system variable

An example for the first round, second round and then third or higher round will be given as to how the ordinary least squares regression model developed by using differences of seasonal averages with a single scoring system variable for a particular round in 2014 tournament was used.

3.3.8.1.1. Ordinary least squares regression model for first round

The ordinary least squares regression model for first round developed by using differences of seasonal averages with a single scoring system variable is:

$\hat{y}$ = (0.78646*Diff in Scoring Margin) + (-1.75276*Diff in Three-Point Field Goals Per Game) + (2.0739*X2 (SINGLE))

University of Connecticut played Prairie View in the first round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.110. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.110. University of Connecticut and Prairie View Statistics

Team	**Score**	**Scoring Margin***	**Three-Point Field Goals***	**Single scoring***
University of Connecticut	87	35.7	7.5	11
Prairie View	44	-4.5	4.4	2
Difference	43	40.2	3.1	9

* Average per game for season

Using the model above, the game between University of Connecticut and Prairie View had a predicted point spread of:

$\hat{y}$ = (0.78646*40.2) + (-1.75276*3.1) + (2.0739*9) = 44.85

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for University of Connecticut, who won the game by a score of 87 to 44.

North Carolina State played BYU in the first round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.111. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.111. North Carolina State and BYU Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Single scoring*
North Carolina State	57	9.6	7.3	0
BYU	72	8.5	6.2	1
Difference	-15	1.1	1.1	-1

* Average per game for season

Using the model above, the game between North Carolina State and BYU had a predicted point spread of:

$\hat{y}$ = (0.78646*1.1) + (- 1.75276*1.1) + (2.0739*-1) = -3.14

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for North Carolina State, who lost the game by a score of 57 to 72.

DePaul played Oklahoma in the first round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.112. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.112. DePaul and Oklahoma Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Single scoring*
DePaul	104	13.2	8.6	3
Oklahoma	100	9	6.4	5
Difference	4	4.2	2.2	-2

* Average per game for season

Using the model above, the game between DePaul and Oklahoma had a predicted point spread of:

$\hat{y}$ = (0.78646*4.2) + (- 1.75276*2.2) + (2.0739*-2) = -4.7

Since $\hat{y}$ <0 this game was coded as an incorrectly predicted loss for DePaul, who won the game by a score of 104 to 100.

Stanford played South Dakota in the first round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.113. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.113. Stanford and South Dakota Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Single scoring*
Stanford	81	17.5	6.5	8
South Dakota	62	2.6	5.7	1
Difference	19	14.9	0.8	7

* Average per game for season

Using the model above, the game between Stanford and South Dakota had a predicted point spread of:

$$\hat{y} = (0.78646*14.9) + (- 1.75276*0.8) + (2.0739*7) = 24.83$$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Stanford, who won the game by a score of 81 to 62.

3.3.8.1.2. Ordinary least squares regression model for second round

The ordinary least squares regression model for second round developed by using differences of seasonal averages with a single scoring system variable is:

$$\hat{y} = (1.3091*\text{Diff in Scoring Margin}) + (-0.33382*\text{Diff in Won-Lost Percentage}) + (0.91812*X2\ (\text{SINGLE}))$$

Oklahoma State played Purdue in the second round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.114. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.114. Oklahoma State and Purdue Statistics

Team	Score	Scoring Margin*	Won-Lost Percentage*	Single scoring*
Oklahoma State	73	11.2	74.2	2
Purdue	66	6.3	72.4	4
Difference	7	4.9	1.8	-2

* Average per game for season

Using the model above, the game between Oklahoma State and Purdue had a predicted point spread of:

$\hat{y}$ = (1.3091*4.9) + (- 0.33382*1.8) + (0.91812*-2) = 3.98

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Oklahoma State, who won the game by a score of 73 to 66.

California played Baylor in the second round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.115. The number of points each of these teams received under the single scoring system for the last two tournaments are found and the difference is taken.

Table 3.115. California and Baylor Statistics

Team	Score	Scoring Margin*	Won-Lost Percentage*	Single scoring*
California	56	4.6	70	7
Baylor	75	22.2	87.9	9
Difference	-19	-17.6	-17.9	-2

* Average per game for season

Using the model above, the game between California and Baylor had a predicted point spread of:

$\hat{y}$ = (1.3091*-17.6) + (- 0.33382*-17.9) + (0.91812*-2) = -18.9

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for California, who lost the game by a score of 56 to 75.

3.3.8.1.3. Ordinary least squares regression model for Third and higher rounds

The ordinary least squares regression model for third and higher rounds developed by using differences of seasonal averages with a single scoring system variable is:

$\hat{y}$ = (1.9762*Diff in Scoring Margin) + (- 0.71222*Diff in Won-Lost Percentage)

Notre Dame played Oklahoma State in the third round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.116.

Table 3.116. Notre Dame and Oklahoma State Statistics

Team	Score	Scoring Margin *	Won-Lost Percentage *
Notre Dame	89	25.6	100
Oklahoma State	72	11.2	74.2
Difference	17	14.4	25.8

* Average per game for season

Using the model above, the game between Notre Dame and Oklahoma State had a predicted point spread of:

$\hat{y}$ = (1.9762*14.4) + (- 0.71222*25.8) = 10.08

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Notre Dame, who won the game by a score of 89 to 72.

Tennessee played Maryland in the third round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.117.

Table 3.117. Tennessee and Maryland Statistics

Team	Score	Scoring Margin *	Won-Lost Percentage *
Tennessee	62	15.5	84.4
Maryland	73	21.3	80
Difference	-11	-5.8	4.4

* Average per game for season

Using the model above, the game between Tennessee and Maryland had a predicted point spread of:

$\hat{y}$ = (1.9762*-5.8) + (- 0.71222*4.4) = -14.6

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for Tennessee, who lost the game by a score of 62 to 73.

3.3.8.2. Examples for seasonal averages models with a double scoring system variable

Results were predicted for every round before the tournament began. Significant differences of seasonal averages for all teams playing in the first round and put into first round model. Values were found for significant differences of seasonal averages for teams predicted to play each other in the second round were placed in second round model and winners of this round were predicted. Values were found for differences of seasonal averages of variables for teams predicted to play each other in the third round were placed in the third round model and winning teams predicted for this round. This process continued until a winner is selected.

The predicted results were then compared against the actual results for each round of the games in the 2014 and 2015 tournaments.

3.3.8.2.1. Examples for seasonal averages models with double scoring system variable

An example will be given as to how the ordinary least squares regression model developed by using differences of seasonal averages with a double scoring system variable for a particular round in the 2014 tournament was used. An example for the first round, second round and then third or higher round is given.

3.3.8.2.1.1. Ordinary least squares regression model for first round

The ordinary least squares regression model for first round developed by using differences of seasonal averages with a double scoring system variable is:

$\hat{Y}$ = (0.73021*Diff in Scoring Margin) + (-1.8507*Diff in Three-Point Field Goals Per Game) + (1.62689*Diff in Blocked shots) + (0.2645*X1 (DOUBLE))

University of Connecticut played Prairie View in the first round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.118. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.118. University of Connecticut and Prairie View Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Blocked shots*	Double scoring*
University of Connecticut	87	35.7	7.5	8.2	94
Prairie View	44	-4.5	4.4	4	2
Difference	43	40.2	3.1	4.2	92

* Average per game for season

Using the model above, the game between University of Connecticut and Prairie View had a predicted point spread of:

$\hat{y}$ = (0.73021*40.2) + (-1.8507*3.1) + (1.62689*4.2) + (0.2645*92) = 54.78

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for University of Connecticut, who won the game by a score of 87 to 44.

North Carolina State played BYU in the first round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.119. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.119. North Carolina State and BYU Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Blocked shots*	Double scoring*
North Carolina State	57	9.6	7.3	2.6	0
BYU	72	8.5	6.2	6	1
Difference	-15	1.1	1.1	-3.4	-1

* Average per game for season

Using the model above, the game between North Carolina State and BYU had a predicted point spread of:

$\hat{y} = (0.73021*1.1) + (-1.8507*1.1) + (1.62689*-3.4) + (0.2645*-1) = -7.03$

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for North Carolina State, who lost the game by a score of 57 to 72.

DePaul played Oklahoma in the first round of the 2014 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.120. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.120. DePaul and Oklahoma Statistics

Team	**Score**	**Scoring Margin***	**Three-Point Field Goals***	**Blocked shots***	**Double scoring***
DePaul	104	13.2	8.6	2.5	4
Oklahoma	100	9	6.4	3.4	10
Difference	4	4.2	2.2	-0.9	-6

* Average per game for season

Using the model above, the game between DePaul and Oklahoma had a predicted point spread of:

$\hat{y} = (0.73021*4.2) + (-1.8507*2.2) + (1.62689*-0.9) + (0.2645*-6) = -4.06$

Since $\hat{y} < 0$ this game was coded as an incorrectly predicted loss for DePaul, who won the game by a score of 104 to 100.

Stanford played South Dakota in the first round of the 2014 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.121. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.121. Stanford and South Dakota Statistics

Team	Score	Scoring Margin*	Three-Point Field Goals*	Blocked shots*	Double scoring*
Stanford	81	17.5	6.5	3.7	38
South Dakota	62	2.6	5.7	2.7	1
Difference	19	14.9	0.8	1	37

* Average per game for season

Using the model above, the game between Stanford and South Dakota had a predicted point spread of:

$$\hat{y} = (0.73021*14.9) + (-1.8507*0.8) + (1.62689*1) + (0.2645*37) = 20.81$$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Stanford, who won the game by a score of 81 to 62.

3.3.8.2.1.2. Ordinary least squares regression model for second round

The ordinary least squares regression model for second round developed by using differences of seasonal averages with a double scoring system variable is:

$$\hat{y} = (1.27219*\text{Diff in Scoring Margin}) + (-0.3248*\text{Diff in Won-Lost Percentage}) + (0.13556*\text{X1 (DOUBLE)})$$

Oklahoma State played Purdue in the second round of the 2014 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.122. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.122. Oklahoma State and Purdue Statistics

Team	Score	Scoring Margin*	Won-Lost Percentage*	Double scoring*
Oklahoma State	73	11.2	74.2	3
Purdue	66	6.3	72.4	6
Difference	7	4.9	1.8	-3

* Average per game for season

Using the model above, the game between Oklahoma State and Purdue had a predicted point spread of:

$\hat{y}$ = (1.27219*4.9) + (- 0.3248*1.8) + (0.13556*-3) = 5.24

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Oklahoma State, who won the game by a score of 73 to 66.

California played Baylor in the second round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.123. The number of points each of these teams received under the double scoring system for the last two tournaments are found and the difference is taken.

Table 3.123. California and Baylor Statistics

Team	**Score**	**Scoring Margin***	**Won-Lost Percentage***	**Double scoring ***
California	56	4.6	70	34
Baylor	75	22.2	87.9	70
Difference	-19	-17.6	-17.9	-36

* Average per game for season

Using the model above, the game between California and Baylor had a predicted point spread of:

$\hat{y}$ = (1.27219*-17.6) + (- 0.3248*-17.9) + (0.13556*-36) = -21.46

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for California, who lost the game by a score of 56 to 75.

3.3.8.2.1.3. Ordinary least squares regression model for third and higher rounds

The ordinary least squares regression model for third and higher rounds developed by using differences of seasonal averages with a double scoring system variable is:

$\hat{y}$ = (1.9762*Diff in Scoring Margin) + (-0.71222*Diff in Won-Lost Percentage)

Notre Dame played Oklahoma State in the third round of the 2014 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.124.

Table 3.124. Notre Dame and Oklahoma State Statistics

Team	**Score**	**Scoring Margin ***	**Won-Lost Percentage ***
Notre Dame	89	25.6	100
Oklahoma State	72	11.2	74.2
Difference	17	14.4	25.8

* Average per game for season

Using the model above, the game between Notre Dame and Oklahoma State had a predicted point spread of:

$\hat{y}$ = (1.9762*14.4) + (- 0.71222*25.8) = 10.08

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Notre Dame, who won the game by a score of 89 to 72.

Tennessee played Maryland in the third round of the 2014 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.125.

Table 3.125. Tennessee and Maryland Statistics

Team	**Score**	**Scoring Margin ***	**Won-Lost Percentage ***
Tennessee	62	15.5	84.4
Maryland	73	21.3	80
Difference	-11	-5.8	4.4

* Average per game for season

Using the model above, the game between Tennessee and Maryland had a predicted point spread of:

$\hat{y}$ = (1.9762*-5.8) + (- 0.71222*4.4) = -14.6

Since $\hat{y} <0$ this game was coded as a correctly predicted loss for Tennessee, who lost the game by a score of 62 to 73.

3.3.9. Results for prediction

3.3.9.1. Results for prediction when using models developed using difference of seasonal averages with a single scoring system variable

In 2014, a continuous process was used to predict the results of the tournament instead of doing round by round validations as in previous section. In other words, a complete bracket was filled out in 2014 before any game was played.

The ordinary least squares regression model for the first round developed by using seasonal averages and a single scoring system variable was used to predict the teams who go to next round. Once the teams in the second round were predicted, the second-round model was used to predict the winners of the second round. This process was continued for the third and higher rounds until the predicted final winner of the game was determined.

A summary of the number of correct and incorrect predictions for each round of the 2014 tournament is given in Table 3.126.

Table 3.126. Prediction Results of each round for 2014: (Ordinary least squares regression model developed by using seasonal averages with a single scoring system variable)

	Correct	Incorrect	Total games
First round	21	11	32
Second round	12	4	16
Third round	7	1	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	1	0	1
	Overall Accuracy		69.84%

A similar process was conducted in predicting 2015 tournament. Namely, a complete bracket was filled out before 2015 tournament started.

A summary of the number of correct and incorrect predictions for each round of the 2015 tournament is given in Table 3.127.

Table 3.127. Prediction Results of each round for 2015: (Ordinary least squares regression model developed by using seasonal averages with a single scoring system variable)

	Correct	Incorrect	Total games
First round	24	12	32
Second round	11	5	16
Third round	6	2	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	1	0	1
	Overall Accuracy		71.43%

3.3.9.2. Results for prediction when using models developed using differences of seasonal averages with a double scoring system variable

A similar process was conducted as in the previous section using the models developed with seasonal averages and a double scoring system variable to predict the results of the 2014 and 2015 tournaments. A complete bracket was filled out for 2014 and 2015 tournaments before any game was played.

The ordinary least squares regression model for the first round developed by using seasonal averages and a double scoring system variable was used to predict the teams who go to next round. Once the teams in the second round were predicted, the second-round model was used to predict the winners of the second round. This process was continued for the third and higher rounds until the predicted final winner of the game was determined.

A summary of the number of correct and incorrect predictions for each round of the 2014 tournament is given in Table 3.128.

Table 3.128. Prediction Results of each round for 2014: (Ordinary least squares regression model developed by using seasonal averages with a double scoring system variable)

	Correct	Incorrect	Total games
First round	24	8	32
Second round	12	4	16
Third round	7	1	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	1	0	1
	Overall Accuracy		74.6%

A similar process was conducted to predict the results for the 2015 tournament. Namely, a complete bracket was filled out before 2015 tournament started. The summary of the number of correct and incorrect predictions for each round of the 2015 tournament is given in Table 3.129.

Table 3.129. Prediction results of each round for 2015: (Ordinary least squares regression model developed by using differences of seasonal averages with a double scoring system variable)

	Correct	**Incorrect**	**Total games**
First round	23	9	32
Second round	10	11	16
Third round	6	2	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	1	0	1
	Overall Accuracy		68.25%

When seasonal averages were used, the models developed by using the double scoring system variable worked better than the ones using the single scoring system variable for the data considered.

3.4. Develop models by using in-game statistics

Data was collected from the results of the NCAA women's basketball tournament of 2014. In-game statistics were collected for 63 games of the 2014 tournament on variables listed in Table 3.2 (Set B). The variables included: Free-Throw Percentage (FT%), Field-Goal

Percentage (FG%),3 Point Goals Percentage (3P%), Offensive Rebounds (OREB), Assists (AST), Steals (ST), Blocks (BLK) and Turnovers (TO). Differences between these variables for the two teams playing each game were found and considered for entry into model.

One ordinary least squares regression model and one logistic regression model were developed by using the data collected from the 2014 season. The first model used ordinary least squares regression with point spread as a response, and the second model used a logistic regression approach with responses recorded as '1' for win and '0' for loss.

3.4.1. Development of ordinary least squares regression model

The ordinary least squares regression model to estimate the point spread based on using significant differences between in-game statistics was found to be:

$$\hat{Y} = (78.00159\text{*Diff in FGP}) + (6.9552\text{*Diff in 3PP}) + (13.57326\text{*Diff in FTP}) + (0.62633\text{*Diff in REB}) + (0.36394\text{*Diff in AST}) + (-1.07784\text{*Diff in TO})$$

The following statistics have positive coefficients associated with them which is to be expected: Difference in FGP, Difference in 3PP, Difference in FTP, Difference in REB and Difference in AST. It is noted that if the team increases Field Goal Percentage by 1% more than other team, the team will on average get 0.78 more points. Each additional rebound over the other team is worth approximately 0.63 points. The only variable that has negative coefficients is Diff in TO. Each additional turnover a team has compared to the opposing team, costs the team an average of 1.08 points over the opposing team.

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 3.130. Table 3.131 gives the steps associated with the stepwise selection technique and Table 3.132 shows the associated R-square values as variables are added to the model. The model with the 6 significant variables explains an estimated 97% of the variation in point spread.

Table 3.130. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
FGP	1	78.00159	7.30785	10.67	<.0001	4.44515
3PP	1	6.95520	4.02554	1.73	0.0894	2.07599
FTP	1	13.57326	2.75585	4.93	<.0001	1.18281
REB	1	0.62633	0.05619	11.15	<.0001	2.67062
AST	1	0.36394	0.10254	3.55	0.0008	2.95907
TO	1	-1.07784	0.10932	-9.86	<.0001	1.29572

Table 3.131. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	FGP		1	0.8063	0.8063	316.563	258.13	<.0001
2	TO		2	0.0722	0.8786	177.760	36.28	<.0001
3	REB		3	0.0691	0.9477	45.0248	79.24	<.0001
4	FTP		4	0.0123	0.9600	22.9930	18.18	<.0001
5	AST		5	0.0102	0.9702	5.1256	19.82	<.0001
6	3PP		6	0.0015	0.9717	4.2330	2.99	0.0894

Table 3.132. Summary of R-squares value

Root MSE	3.73453	R-Square	0.9717
Dependent Mean	-5.00000	Adj R-Sq	0.9687
Coeff Var	-74.69057		

3.4.2. Development of logistic regression model

A logistic regression model to help estimate the probability of the team of interest winning the game was developed and found to be:

$$\pi\text{REB, FGP, FTP}=\frac{e^{0.3031REB+32.0237FGP+13.5347FTP}}{1+e^{0.3031REB+32.0237FGP+13.5347FTP}}$$

Where π (REB, FGP,FTP) is the estimated probability that the team of interest will win the game with differences of in-game statistics in rebounds, difference of in-game statistics in Field Goal Percentage and difference of in-game statistics in Free Throw Percentage in model.

Table 3.133 shows the steps for the stepwise selection technique and Table 3.134 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 3.135 shows the Hosmer and Lemeshow Test [8] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.992 indicating that there was no evidence to reject using the logistic regression model.

Table 3.133. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	REB		1	1	36.3735		<.0001
2	FGP		1	2	12.1608		0.0005
3	TO		1	3	6.0411		0.0140
4	FTP		1	4	3.6833		0.0550
5		TO	1	3		1.4325	0.2314
6	TO		1	4	6.6876		0.0097
7		TO	1	3		1.4325	0.2314

Table 3.134. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
FGP	1	32.0237	14.2846	5.0258	0.0250
FTP	1	13.5347	7.9615	2.8900	0.0891
REB	1	0.3031	0.1305	5.3937	0.0202

Table 3.135. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
1.5360	8	0.9921

3.4.3. Validating models

3.4.3.1. Validating first round using models developed

The ordinary least squares regression model developed by using in-game statistics was used to predict the results of the 2015. The logistic regression model was also used to predict the results of the 2015 tournament. It is noted that the 2015 season was not used in the development of the models.

Table 3.136 gives the results as to how accurately the ordinary least squares regression model developed by using in-game statistics predicted the winning teams of the first round of the NCAA 2015 women's basketball game.

Table 3.136. Accuracy of ordinary least squares regression model developed by in-game statistics when validating first round of 2015

Point spread		**predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	26	0	26
	Loss	1	5	6
	Total	27	5	32
Overall Accuracy				96.88%

Table 3.137 gives the results as to how accurately the logistic regression model developed by using in-game statistics when predicting the winning teams of the first round of the NCAA 2015 women's basketball game.

Table 3.137. Accuracy of logistic regression model developed by in-game statistics when validating first round of 2015

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	23	3	26
	Loss	2	4	6
	Total	25	7	32
Overall Accuracy				84.38%

3.4.3.2. Validating second round using models developed

Table 3.138 gives the results as to how accurately the ordinary least squares regression model developed by using in-game statistics predicts the second round of the NCAA 2015 women's basketball game.

Table 3.138. Accuracy of ordinary least squares regression model developed by in-game statistics when validating second round of 2015

Point spread		**predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	7	0	7
	Loss	0	9	9
	Total	7	9	16
Overall Accuracy				100%

Table 3.139 gives the results as to how accurately the logistic regression model developed by using in-game statistics was in predicting the second round of the NCAA 2015 women's basketball game.

Table 3.139. Accuracy of logistic regression model developed by in-game statistics when validating second round of 2015

Logistic		**predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	7	0	7
	Loss	1	8	9
	Total	8	8	16
Overall Accuracy				93.75%

3.4.3.3. Validating third and higher rounds using models developed

Table 3.140 gives the results as to how accurately the ordinary least squares regression model developed by using in-game statistics was in predicting the third and higher rounds of the NCAA 2015 women's basketball game.

Table 3.140. Accuracy of ordinary least squares regression model developed by in-game statistics when validating third and higher rounds of 2015

Point spread		predicted		
		Win	**Loss**	**Total**
Actual	**Win**	11	0	11
	Loss	1	3	14
	Total	12	3	15
Overall Accuracy				93.33%

Table 3.141 gives the results as to how accurately the logistic regression model developed by using in-game statistics was in predicting the third and higher rounds of the NCAA 2015 women's basketball game.

Table 3.141. Accuracy of logistic regression model developed by in-game statistics when validating third and higher rounds of 2015

Logistic		Predicted		
		Win	**Loss**	**Total**
Actual	**Win**	9	0	9
	Loss	2	4	6
	Total	11	4	15
Overall Accuracy				86.67%

3.4.4. Bracketing the 2016 tournament before tournament begins - Prediction

Since the in-game statistics will not be available before the tournaments began, significant differences of in-game statistics were replaced with seasonal averages for the current year of the associated statistics. The seasonal averages of these statistics for all teams playing in the 2016 tournament were collected.

Differences of seasonal averages for teams playing each other in the first round were placed into the model for each game in the first round and the winning teams were predicted.

Namely, if $\hat{y}$ is great than 0, a predicted win for the team of interest was coded. If $\hat{y}$ is less than 0, a predicted loss for the team of interest was coded.

To verify the accuracy of prediction results for the logistic regression model for the first round, a similar process was conducted. Differences of seasonal averages were placed into the logistic model instead of differences of in-game statistics. If π xi is greater than 0.5, a predicted win was coded. If π xi is less than 0.5, a predicted loss was coded for the team of interest.

Once the teams making it to the second round were predicted, the same model was used to predict the winners of the second round. This process continued for the third and higher rounds.

In 2016, a continuous process was used in predicting winners of all games instead of doing round by round predictions as in 2015 using both the ordinary least squares and logistic models. Namely, a complete bracket was filled out in 2016 before any game was played.

3.4.4.1. Example for in-game statistics models when predicting 2016

An example will be given as to how the ordinary least squares regression model developed by using in-game statistics for a particular round was used for each round in 2016 tournament.

3.4.4.1.1. Ordinary least squares regression models

3.4.4.1.1.1. Ordinary least squares regression model for first and higher rounds

The ordinary least squares regression model to estimate the point spread based on using difference between in-game statistics of the significant variables was found to be the following:

$$\hat{y} = (78.00159*\text{Diff in FGP}) + (6.9552*\text{Diff in 3PP}) + (13.57326*\text{Diff in FTP}) + (0.62633*\text{Diff in REB}) + (0.36394*\text{Diff in AST}) + (-1.07784*\text{Diff in TO})$$

Seton Hall played Duquesne in the first round of the 2016 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.142.

Table 3.142. Seton Hall and Duquesne Statistics

Team	Score	FGP*	3PP*	FTP*	REB*	AST*	TO*
Seton Hall	76	0.4247392	0.33232628	0.71061644	39.688	11.938	13.875
Duquesne	97	0.40751043	0.33550914	0.7357513	42.324	16.412	14.38235
Difference	-21	0.01722876	-0.00318285	-0.02513486	-2.636	-4.474	-0.5073529

* Average per game for season

Using the model above, the game between Seton Hall and Duquesne had a predicted point spread of:

$\hat{y}$ = (78.00159*0.01722876) + (6.9552*-0.00318285) + (13.57326*-0.02513486) + (0.62633*-2.636) + (0.36394*-4.474) + (-1.07784*-0.5073529) = -1.75

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for Seton Hall, who lost the game by a score of 76 to 97.

South Florida played Colorado State in the first round of the 2016 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.143.

Table 3.143. South Florida and Colorado State Statistics

Team	Score	FGP*	3PP*	FTP*	REB*	AST*	TO*
South Florida	48	0.41004184	0.34410339	0.78322785	42.794	13	12.67647
Colorado	45	0.44444444	0.34385382	0.70804598	37.515	15	12.0303
Difference	3	-0.0344026	0.00024957	0.07518187	5.279	-2	0.6461676

* Average per game for season

Using the model above, the game between South Florida and Colorado had a predicted point spread of:

$\hat{y}$ = (78.00159*-0.0344026) + (6.9552*0.00024957) + (13.57326*0.07518187) + (0.62633*5.279) + (0.36394*-2) + (-1.07784*0.6461676) = 0.65

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for South Florida, who lost the game by a score of 48 to 45.

Louisville played Central Arkansas in the first round of the 2016 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.144.

Table 3.144. Louisville and Central Arkansas Statistics

Team	Score	FGP*	3PP*	FTP*	REB*	AST*	TO*
Louisville	87	0.43536761	0.32705479	0.71473851	38.706	16.118	15.05882
Central Arkansas	60	0.42322725	0.34878049	0.73483536	38.969	14.188	14.15625
Difference	27	0.0121404	-0.021726	-0.020097	-0.263	1.93	0.902574

* Average per game for season

Using the model above, the game between Louisville and Central Arkansas had a predicted point spread of:

$\hat{y}$ = (78.00159*0.0121404) + (6.9552*-0.021726) + (13.57326*-0.020097) + (0.62633*-0.263) + (0.36394*1.93) + (-1.07784*0.902574) = 0.09

Since $\hat{y}$ >0 this game was coded as an incorrectly predicted win for Louisville, who won the game by a score of 87 to 60.

Miami (Florida) played South Dakota State in the first round of the 2016 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.145.

Table 3.145. Miami (Florida) and South Dakota State Statistics

Team	Score	FGP*	3PP*	FTP*	REB*	AST*	TO*
Miami (Florida)	71	0.43057571	0.33695652	0.63584906	39.697	15.576	15.57576
South Dakota State	74	0.41496921	0.34635417	0.69391635	39.853	14.618	12.94118
Difference	-3	0.0156065	-0.009398	-0.058067	-0.156	0.958	2.634581

* Average per game for season

Using the model above, the game between Miami (Florida) and South Dakota State had a predicted point spread of:

$\hat{y}$ = (78.00159*0.0156065) + (6.9552*-0.009398) + (13.57326*-0.058067) + (0.62633*-0.156) + (0.36394*0.958) + (-1.07784*2.634581) = -2.22

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for Miami (Florida), who lost the game by a score of 71 to 74.

3.4.4.1.2. Logistic regression models

3.4.4.1.2.1. Logistic regression model for first and higher rounds

The logistic regression model to help estimate the probability of the team of interest winning based on significant differences of in-game statistics was found to be:

$$\pi\text{REB, FGP, FTP} = \frac{e^{0.3031REB+32.0237FGP+13.5347FTP}}{1+e^{0.3031REB+32.0237FGP+13.5347FTP}}$$

Seton Hall played Duquesne in the first round of the 2016 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.146.

Table 3.146. Seton Hall and Duquesne Statistics

Team	Score	REB*	FGP*	FTP*
Seton Hall	76	39.688	0.4247392	0.71061644
Duquesne	97	42.324	0.40751043	0.7357513
Difference	-21	-2.636	0.01722876	-0.02513486

* Average per game for season

Using the model above, the game between Seton Hall and Duquesne had an estimated probability of winning the game of:

$$\pi\ (-2.636, 0.017, -0.025) = \frac{e^{0.3031*-2.636+32.0237*0.017+13.5347*-0.025}}{1+e^{0.3031*-2.636+32.0237*0.017+13.5347*-0.025}} = 0.36$$

Since $\pi < 0.5$ this game was coded as a correctly predicted loss for Seton Hall, who lost the game by a score of 76 to 97.

BYU played Missouri in the first round of the 2016 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.147.

Table 3.147. BYU and Missouri Statistics

Team	Score	REB*	FGP*	FTP*
BYU	69	37.485	0.42667375	0.70289855
Missouri	78	38.781	0.43348624	0.76256499
Difference	-9	-1.296	-0.00681249	-0.05966644

* Average per game for season

Using the model above, the game between BYU and Missouri had an estimated probability of winning the game of:

$$\pi\ (-1.296, -0.007, -0.06) = \frac{e^{0.3031*-1.296+32.0237*-0.007+13.5347*-0.06}}{1+e^{0.3031*-1.296+32.0237*-0.007+13.5347*-0.06}} = 0.19$$

Since $\pi < 0.5$ this game was coded as a correctly predicted loss for BYU, who lost the game by a score of 69 to 78.

Louisville played Central Arkansas in the first round of the 2016 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.148.

Table 3.148. Louisville and Central Arkansas Statistics

Team	**Score**	**REB***	**FGP***	**FTP***
Louisville	87	38.706	0.43536761	0.71473851
Central Arkansas	60	38.969	0.42322725	0.73483536
Difference	27	-0.263	0.0121404	-0.020097

* Average per game for season

Using the model above, the game between Louisville and Central Arkansas had an estimated probability of winning the game of:

$$\pi\ (-0.263,0.012,-0.02) = \frac{e^{0.3031*-0.263+32.0237*0.012+13.5347*-0.02}}{1+e^{0.3031*-0.263+32.0237*0.012+13.5347*-0.02}} = 0.51$$

Since $\pi > 0.5$ this game was coded as a correctly predicted win for Louisville, who won the game by a score of 87 to 60.

Miami (Florida) played South Dakota State in the first round of the 2016 Tournament. Data on significant differences of seasonal averages was collected and displayed in Table 3.149.

Table 3.149. Miami (Florida) and South Dakota State Statistics

Team	**Score**	**REB***	**FGP***	**FTP***
Miami (Florida)	71	39.697	0.43057571	0.63584906
South Dakota State	74	39.853	0.41496921	0.69391635
Difference	-3	-0.156	0.0156065	-0.058067

* Average per game for season

Using the model above, the game between Miami (Florida) and South Dakota State had an estimated probability of winning the game of:

$$\pi\ (-0.156,0.016,-0.058) = \frac{e^{0.3031*-0.156+32.0237*0.016+13.5347*-0.058}}{1+e^{0.3031*-0.156+32.0237*0.016+13.5347*-0.058}} = 0.42$$

Since $\pi < 0.5$ this game was coded as a correctly predicted loss for Miami (Florida), who lost the game by a score of 71 to 74.

3.4.5. Results for prediction by using in-game statistics models

In 2016, a continuous process was used in verifying the models instead of doing round by round predictions as in the previous section. In other words, a complete bracket was filled out in 2016 before any game was played.

One ordinary least square regression model and one logistic regression model were used to predict each round of NCAA women's basketball tournament of 2016. Since the in-game statistics would not be available before the tournaments began, seasonal averages were entered into in-game model to predict the winner of the basketball game for 2016.

A summary of the number of correct and incorrect predictions for ordinary least squares regression model for each round of the 2016 tournament is given in Table 3.150.

Table 3.150. Prediction results of each round for 2016: (Ordinary least squares regression model developed by in-game statistics)

	Correct	**Incorrect**	**Total games**
First round	23	9	32
Second round	9	7	16
Third round	2	6	8
Fourth round	1	3	4
Fifth round	1	1	2
Final round	1	0	1
	Overall Accuracy		58.73%

A summary of the number of correct and incorrect predictions for logistic regression model for each round of the 2016 tournament is given in Table 3.151.

Table 3.151. Prediction results of each round for 2016: (Logistic regression model developed by in-game statistics)

	Correct	**Incorrect**	**Total games**
First round	28	4	32
Second round	9	7	16
Third round	3	5	8
Fourth round	1	3	4
Fifth round	1	1	2
Final round	1	0	1
	Overall Accuracy		68.25%

It is noted that both models correctly predicted the winning team.

3.5. Conclusion

3.5.1. Validation - Models developed by using seasonal averages

To verify the accuracy of prediction results for the ordinary least squares regression models developed for each round using differences of the seasonal averages, data from the 2014 tournament was used. The ordinary least squares regression model developed by using differences in ranks of seasonal averages with either a single or double scoring system variable for the first round had approximately a 62.63% and a 78.13% chance of correctly predicting the results, respectively. The ordinary least squares regression model developed by using seasonal averages with either a single or double scoring system variable for the first round had approximately a 65.63% chance of correctly predicting the results.

The ordinary least squares regression model developed by using differences in ranks of seasonal averages with either a single or double scoring system variable for the second round had approximately a 75% chance of correctly predicting the results. The ordinary least squares regression model developed by using seasonal averages with either a single or double scoring system variable for the second round had approximately a 75% chance of correctly predicting the results.

The ordinary least squares regression model developed by using differences in ranks of seasonal averages with either a single or double scoring system variable for the third and higher rounds had approximately a 93.33% chance of correctly predicting the results. The ordinary least squares regression model developed by using seasonal averages with either a single or double scoring system variable for the third and higher rounds had approximately a 80% chance of correctly predicting the results.

3.5.2. Prediction - Models developed by using seasonal averages

In 2015, a continuous process was used to predict the winning team in each round before the tournament started instead of doing round by round predictions as in 2014. Namely, a complete bracket was filled out in 2015 before any game was played. When the differences of the seasonal averages for both teams for all previously mentioned variables were considered for entry in the ordinary least squares models which developed by using differences in ranks of seasonal averages with either a single or double scoring system variable, the models had approximately a 74.6% and 73.02% chance of correctly predicting the winner of a basketball game, respectively. When the differences of the seasonal averages for both teams for all previously mentioned variables were considered for entry in the ordinary least squares models which developed by using seasonal averages with either a single or double scoring system variable, the models had approximately a 71.43% chance of correctly predicting the winner of a basketball game.

3.5.3. Validation - Model developed by using in-game statistics

To verify the accuracy of prediction results for the ordinary least squares regression model developed by using in-game statistics, differences of the in-game statistics for both teams for all previously mentioned variables listed in Table 3.2 (Set B) were placed in the model. The ordinary least squares regression model and the logistic regression model for the first round had approximately a 96.88% and 84.38% chance of correctly predicting the results, respectively. The ordinary least squares regression model and the logistic regression model for the second round had approximately a 100% and 93.75% chance of correctly predicting the results, respectively. The ordinary least squares regression model and the logistic regression model for the third and higher rounds had approximately a 93.33% and 86.67% chance of correctly predicting the results, respectively.

3.5.4. Prediction - Model developed by using in-game statistics

When the differences of the seasonal averages were placed into the model developed by using differences of in-game statistics, the ordinary least squares regression model and logistic regression model correctly predicted 59% and 68%, respectively, of the games correctly.

It is noted that the predictions were done and brackets filled out before the tournament began. The accuracy is lower because teams predicted to play in the second round or higher round might not have actually made it to those rounds.

3.5.5. Overall comparisons

Both the ordinary least squares regression model and the logistic regression model developed by using in-game statistics work well when the in-game statistics are known.

When predicting results for future tournaments without in-game statistics given, the ordinary least squares regression model has an overall accuracy of 59% and logistic regression model has an overall accuracy of 68% chance of correctly pick the winner of each game in NCAA women's basketball tournament. This result was not surprising since the models were developed by using in-game statistics and replaced with seasonal averages when doing the prediction.

Overall, ordinary least squares models developed by using seasonal averages had an overall accuracy is 75% works slightly better than models developed by using in-game statistics when estimating the point spread of a NCAA women's basketball tournament game.

3.6. References

[1] NCAA Division I Women's Basketball Tournament. Retrieved October 10, 2017, from https://en.wikipedia.org/wiki/NCAA_Division_I_Women%27s_Basketball_Tournament.

[2] Road to the Championship. Retrieved October 10, 2017, from http://www.ncaa.com/womens-final-four/road-to-the-championship.

[3] Ranking Summary. Retrieved October 10, 2015, from http://web1.ncaa.org/stats/StatsSrv/ranksummary.

[4] Women's Basketball. Retrieved October 10, 2017, from http://www.ncaa.com/stats/basketball-women/d1.

[5] Basketball Glossary of Statistics. Retrieved October 10, 2017, from http://www.hometeamsonline.com/teams/popups/Glossary.asp?s=basketball.

[6] Women's Basketball 2015-2016 and 2016-2017 Rules. Retrieved October 10, 2017, from http://www.ncaapublications.com/productdownloads/WBR17.pdf Web.

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

[8] Hosmer, David W.; Lemeshow, Stanley (2013). *Applied Logistic Regression.* New York: Wiley. ISBN 978-0-470-58247-3

CHAPTER 4. BRACKETING NCAA WOMEN'S VOLLEYBALL TOURNAMENT

4.1. Introduction

4.1.1. The history of NCAA women's volleyball tournament

The NCAA division I women's volleyball tournament is the annual championship in women's volleyball from teams in division I contested by the National Collegiate Athletic Association(NCAA) each winter since 1981. Volleyball was added to the NCAA championship program for the 1981-1982 school year. There were only 20 schools competing for the first NCAA championship which held in 1981. The tournament expanded gradually, and its current size of 64 teams was attained in 1998 (NCAA - Volleyball [1]).

4.1.2. The playing rule and structure

There are 330 NCAA member institutions that sponsor division I women's volleyball teams and are eligible to compete in the National Championship. There are 64 teams that play 32 games to compete in a single elimination tournament for the first round of the NCAA division I women's volleyball tournament championship. Of the 64 teams, 32 teams will receive automatic qualification while the rest 32 teams are selected by the division I women's volleyball committee (Road to the Championship [2]).

For the first round, there will be 64 teams competing in single-elimination to advance to second round. The 32 advancing teams then compete against each other in single-elimination second round competition. The winning teams will advance to the regional round. For the regional round, there will be 16 teams competing in single-elimination regional semifinal competition. The advancing teams then compete against each other in the single-elimination regional final. The winning team for the four regions will advance to the NCAA women's volleyball championship final game. There will be 4 teams competing in single-elimination semifinal and the advancing teams then compete against each other in the national championship title (Road to the Championship [2]). Figure 2 shows the 2015 - 2016 NCAA women's volleyball tournament bracket.

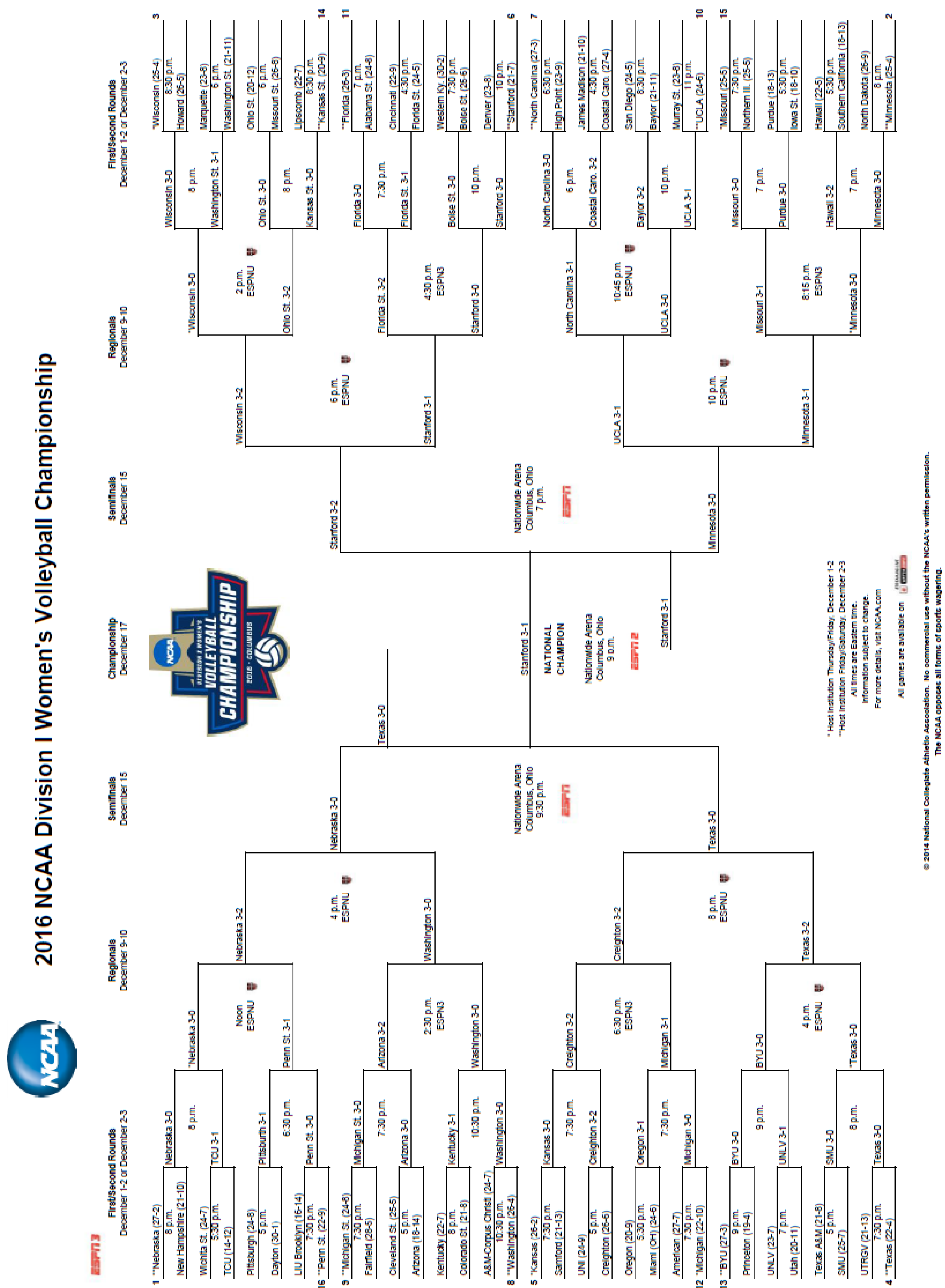

Figure 2. The NCAA women's volleyball tournament bracket for the 2015 – 2016 season. (This bracket is downloaded from: http://www.ncaa.com/interactive-bracket/volleyball-women/d1)

4.1.3. The research objectives for this study

The research objectives for this study are as follows:

1) Develop ordinary least squares regression models for Round 1, Round 2 and Rounds 3-6 with point spread being the dependent variable and using differences in ranks of seasonal averages of various variables, to predict winners of volleyball games in each of those rounds for the NCAA women's volleyball tournament; and

2) Develop logistic regression models for Round 1, Round 2 and Rounds 3-6 that estimate the probability of a team winning the game by using differences in ranks of seasonal averages, to predict winners of volleyball games in each of those rounds for the NCAA women's volleyball tournament; and

3) Develop ordinary least squares regression models for Round 1, Round 2 and Rounds 3-6 with point spread being the dependent variable by using difference of seasonal averages of various variables, to predict winners of volleyball games in each of those rounds for the NCAA women's volleyball tournament; and

4) Develop logistic regression models for Round 1, Round 2 and Rounds 3-6 that estimate the probability of a team winning the game by using difference of seasonal averages, to predict winners of volleyball games in each of those rounds for the NCAA women's volleyball tournament; and

5) Develop one ordinary least squares regression model by using in-game statistics, to explain the variation of the point spread of a women's volleyball game and then use this model to predict the winners of the volleyball games for the NCAA women's volleyball tournament by estimating the significant in-game statistics with differences in seasonal averages of the statistics between the two teams playing; and

6) Develop one logistic regression model by using in-game statistics, and use this model to predict winners by replacing the significant in-game statistics with the differences in seasonal averages of the statistics for the two teams playing.

In order to accomplish objectives 1 and 2, data was collected for three years of the NCAA women's volleyball tournament. This included data from the 2011, 2012 and 2013 tournaments. Differences in ranks of seasonal averages were collected for all the teams in the 2011 tournament on the following variables in Table 4.1 (Set A). Differences in ranks of seasonal averages were also collected on the same variables for all teams playing against each other in the 2012 and 2013 tournament. The developed models are given in Section 4.2.

Table 4.1. Set A - Variables in consideration for seasonal average

Variables in consideration	**Definitions**
Aces Per Set	A serve that results directly in a point when a player attempts to serve the ball over the net into the opponent's court for each set. [3]
Assists Per Set	When a player passer, sets or digs ball to teammate who gets a kill for each set. [3]
Blocks Per Set	Player(s) block leads directly to a point for each set. [3]
Digs Per Set	When a player receives an attacked ball and keeps the ball in play for each set. [3]
Hitting Percentage	Hitting Percentage = (Total kills – Total Errors)/ Total Attempts. [3]
Kills Per Set	An attack that directly leads to a point for each set. [3]
Match W-L Percentage	Match W-L Percentage = Numbers of games won / Total sets played. [3] Note: The value for Match W-L Percentage will between 0 to 1.

In order to accomplish objectives 3 and 4, data was collected for three years of the NCAA women's volleyball tournament. This included data from the 2011, 2012 and 2013 tournaments. Seasonal averages were collected for all the teams playing each other in the 2011 tournament on the same variables listed in Table 4.1 (Set A). Seasonal averages were also collected on the same variables for all teams playing each other in the 2012 and 2013 tournament. The developed models are given in Section 4.3.

For research objectives 5 and 6, data was collected from the NCAA women's volleyball tournament of 2015. In-game statistics were collected for 37 games of 63 games of the 2015 tournament on the variables listed in Table 4.2 (Set B): Attack Kill, Attack Error, Attack Percentage, SERVE SA, SRV RE, Digs and Blocks. The developed models are given in Section 4.4.

Table 4.2. Set B - Variables in consideration for in-game statistics

Variables in consideration	**Definitions**
Attack Kill	An attack that directly leads to a point. [3]
Attack Error	An attack that directly results in a point for the opposing team. [3]
Attack Percentage	Attack Percentage = (Total kills – Total Errors)/ Total Attempts. [3]
SERVE SA (Service ace)	A service ace (SA) is a serve that results directly in a point when a player attempts to serve the ball over the net into the opponent's court. [3]
SRV RE (Reception Error)	When a result for a point for the opposing team a player of team must be charged with a reception error. [3]
Digs	When a player receives an attacked ball and keeps the ball in play. [3]
Blocks	Player(s) block leads directly to a point. [3]

4.2. Model developed by using differences in ranks of seasonal averages

4.2.1. Develop models by using differences in ranks of seasonal averages

All data was collected from NCAA.COM [4]. Data for the ranks of the seasonal averages of the variables of interest were collected before the tournament started. For example, the first game of NCAA 2011 women's volleyball tournament was held on December 1, 2011. The ranks of the seasonal averages for each of the variables were based on games played through November 27, 2011.

Data was collected for three years of the NCAA women's volleyball tournament. This included 2011, 2012 and 2013 tournaments. The ranks for the seasonal averages for the variables of interest for each team were collected for all the teams in the 2011 tournament on the variables listed in Table 4.1 (Set A). The variables included: Aces Per Set, Assists Per Set, Blocks Per Set, Digs Per Set, Hitting Percentage, Kills Per Set and Match W-L Percentage. Ranks of the seasonal averages for the variables of interest for each team were also collected for all teams playing in the 2012 and 2013 tournaments.

4.2.2. Develop models for the first round using differences in ranks of seasonal averages

4.2.2.1. Develop ordinary least squares regression models

The response variable for the ordinary least squares regression model was the point spread of the game in the order of the team of interest minus the opposing team. A positive point spread indicates a win for the team of interest and a negative value indicates a loss for the team of interest. There were 192 teams playing 96 games in first rounds of the tournaments in 2011, 2012 and 2013. For the half games of the first round in the three years, the team of interest is the stronger team (higher seed numbers), the point spread was obtained by using the stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the remainder of games in the first round, the team of interest is the weaker team (lower seed numbers), the point spread was acquired by using the score of weaker team (lower seed number) minus the stronger team (higher seed number).

The intercept was excluded when developing the models because the models should give the same results regardless of the ordering of the teams in the model. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences of the ranks of seasonal averages for all the variables previously given in Table 4.1 (Set A) between the two teams were considered for entry in the model in the order team of interest minus opposing team.

4.2.2.1.1. Development of ordinary least squares regression model for first round

The ordinary least squares regression model to predict the winning team for each game in the first round based on using difference between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-0.01228 * \text{Diff\_Assists}) + (-0.00925 * \text{Diff\_Blocks})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.3. Table 4.4 gives the steps associated with the stepwise selection technique and Table 4.5 shows the associated R-square values as variables are added to the model. The model with the 2 significant variables explains an estimated 38% of the variation in point spread.

Table 4.3. Point spread model parameter estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Diff_Assists	1	-0.01228	0.00253	-4.85	<.0001	1.00941
Diff_Blocks	1	-0.00925	0.00171	-5.42	<.0001	1.00941

Table 4.4. Summary of stepwise selection for point spread model

Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Diff_Hitting%_		1	0.2654	0.2654	19.7568	34.33	<.0001
2	Diff_Blocks		2	0.0773	0.3427	9.7928	11.05	0.0013
3	Diff_Assists		3	0.0507	0.3934	3.9388	7.78	0.0064
4		Diff_Hitting%_	2	0.0097	0.3837	3.4442	1.49	0.2252

Table 4.5. Summary of R-squares value

Root MSE	2.00969	R-Square	0.3837
Dependent Mean	-0.14583	Adj R-Sq	0.3706
Coeff Var	-1378.07576		

4.2.2.2. Develop logistic regression models for first round

The logistic regression model was also fit to the data with the dependent variable recorded as '1' for win and '0' for loss for the team of interest. This model estimates the probability of a win for the team of interest. Team of interest was the stronger team (higher seed numbers) in half of the games and the weaker team (lower seed numbers) in the other half of the games.

No intercept was used during the development of the logistic regression model since the ordering of the teams in the model should not matter. Stepwise selection was used with an α value of 0.1 for both entry and exit when determining the significant variables in developing the logistic regression model. The differences in ranks of the seasonal averages for both teams for all previously mentioned variables listed in Table 4.1 (Set A) were considered for entry in the model.

4.2.2.2.1. Development of logistic regression model for the first round

The logistic regression model to predict the winning team for each game in the first round was developed and found to be:

$$\pi(\text{Diff\_Blocks, Diff\_Kills})=\frac{e^{-0.0114*\text{Diff\_Blocks}-0.0169*\text{Diff\_Kills}}}{1+e^{-0.0114*\text{Diff\_Blocks}-0.0169*\text{Diff\_Kills}}}$$

Where π (Diff_Blocks, Diff_Kills) is the estimated probability that the team of interest will win the game with differences in ranks of seasonal averages in blocks and differences in ranks of seasonal averages in kills in model.

Table 4.6 shows the steps for the stepwise selection technique and Table 4.7 gives the parameter estimates, the standard errors and associated p-values when both the variables are in the model. Table 4.8 shows the Hosmer and Lemeshow Test [5] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.1546 indicating that there was no evidence to reject using the logistic regression model.

Table 4.6. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	Diff_Blocks		1	1	19.9778		<.0001
2	Diff_Kills		1	2	12.6898		0.0004

Table 4.7. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Diff_Blocks	1	-0.0114	0.00285	16.0866	<.0001
Diff_Kills	1	-0.0169	0.00510	11.0503	0.0009

Table 4.8. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
11.9243	8	0.1546

4.2.3. Develop models for the second round using differences in ranks of seasonal averages

4.2.3.1. Develop ordinary least squares regression models for second round

There were 96 teams playing 48 games in second rounds of the tournaments in 2011 to 2013. For the first half of the second round, the point spread was obtained by using the stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the weaker team (lower seed numbers) minus the stronger team (higher seed numbers). No intercept was used when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences in ranks of the seasonal averages of the variables listed in Table 4.1 (Set A) between the two teams were considered for entry in the model.

4.2.3.1.1. Development of ordinary least squares regression model for the second round

The ordinary least squares regression model to predict the winning team for each game in the second round based on using differences in ranks of seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = 0.00834 * \text{Diff\_Digs}$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.9. Table 4.10 gives the steps associated with the stepwise selection technique and Table 4.11 shows the associated R-square values as variables are added to the model. The model with the 1 significant variable explains an estimated only 15% of the variation in point spread.

Table 4.9. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Diff_Digs	1	0.00834	0.00292	2.85	0.0064	1.00000

Table 4.10. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Diff_Digs		1	0.1478	0.1478	0.5467	8.15	0.0064

Table 4.11. Summary of R-squares value

Root MSE	2.23304	R-Square	0.1478
Dependent Mean	-0.06250	Adj R-Sq	0.1296
Coeff Var	-3572.86064		

4.2.3.2. Develop logistic regression models for the second round

The logistic regression model was also fit for the data with responses recorded as'1' for win and '0' for loss for the team of interest. With the model estimating the probability of a win for the team of interest. No intercept was used during the development of the logistic regression model. Stepwise selection was used with an α value of 0.1 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences in ranks of the seasonal averages of all previously mentioned variables in Table 4.1 (Set A) for the two teams playing in each game were considered for entry in the model.

4.2.3.2.1. Development of logistic regression model for the second round

A logistic regression model to predict the winning team for each game in the second round was developed and found to be:

$$\pi(\text{Diff\_Digs})=\frac{e^{0.00653*\text{Diff\_Digs}}}{1+e^{0.00653*\text{Diff\_Digs}}}$$

Where π (Diff_Digs) is the estimated probability that the team of interest will win the game with differences in ranks of seasonal averages in digs in model.

Table 4.12 shows the steps for the stepwise selection technique and Table 4.13 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 4.14 shows the Hosmer and Lemeshow Test [5] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.427 indicating that there was no evidence to reject using the logistic regression model.

Table 4.12. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq	Variable
	Entered	Removed		In	Chi-Square	Chi-Square		Label
1	Diff_Digs		1	1	5.0762		0.0243	Diff_Digs

Table 4.13. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Diff_Digs	1	0.00653	0.00304	4.5953	0.0321

Table 4.14. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
8.0668	8	0.4270

4.2.4. Develop models for the third and higher rounds using differences in ranks of seasonal averages

4.2.4.1. Develop ordinary least squares regression model

There were 90 teams playing 45 games in third and higher rounds of the tournaments in 2011 to 2013. For the first 24 games of the second round, the point spread was obtained by using the stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the scores of weaker team (lower seed numbers) minus the stronger team (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences in ranks of the seasonal averages of the previously mentioned variables listed in Table 4.1 (Set A) between the two teams were considered for entry in the model.

4.2.4.1.1. Development of ordinary least squares regression model for the third and higher rounds

The ordinary least squares regression model to predict the winning team for each game in the third and higher rounds based on using differences in ranks of seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = -0.02617^{*}\ \text{Diff\_Match\_W-L\%}$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.15. Table 4.16 gives the steps associated with the stepwise selection technique and Table 4.17 shows the associated R-square values as variables are added to the model. The model with the only 1 significant variable explains an estimated 17% of the variation in point spread.

Table 4.15. Point spread model parameter estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Diff_Match_W-L%_	1	-0.02617	0.00871	-3.01	0.0044	1.00000

Table 4.16. Summary of stepwise selection for point spread model

Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Diff_Match_W-L%_		1	0.1704	0.1704	3.5610	9.04	0.0044

Table 4.17. Summary of R-squares value

Root MSE	2.16675	R-Square	0.1704
Dependent Mean	0.11111	Adj R-Sq	0.1515
Coeff Var	1950.07838		

4.2.4.2. Develop logistic regression model for the third and higher rounds

The logistic regression model was also fit for the data with responses recorded as‘1’ for win and ‘0’ for loss for the team of interest. No intercept was used during the development of the logistic regression model. Stepwise selection was used with an α value of 0.1 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences in ranks of the seasonal averages of all previously mentioned variables in Table 4.1 (Set A) between the two teams were considered for entry in the model.

4.2.4.2.1. Development of logistic regression model for the third and higher rounds

A logistic regression model to help predict the winning team for each game in the third and higher rounds was developed and found to be:

$$\pi(\text{Diff\_W-L\%}) = \frac{e^{-0.0295*\text{Diff\_W-L\%}}}{1+e^{-0.0295*\text{Diff\_W-L\%}}}$$

Where π (Diff_W-L%) is the estimated probability that the team of interest will win the game with differences in ranks of seasonal averages in won-lost percentage in model.

Table 4.18 shows the steps for the stepwise selection technique and Table 4.19 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 4.20 shows the Hosmer and Lemeshow Test [5] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.1521 indicating that there was no evidence to reject using the logistic regression model.

Table 4.18. Summary of stepwise selection for logistic regression model

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Diff_W-L%		1	1	7.5986		0.0058	Diff_Match_W-L%_

Table 4.19. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Diff_W-L%	1	-0.0295	0.0121	5.9027	0.0151

Table 4.20. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
10.7029	7	0.1521

4.2.5. Validating first round using models developed

4.2.5.1. Verification of the models developed by using differences in ranks of seasonal averages

Using the ordinary least squares regression model developed for the first round, the point spread of the 32 games in the first round of the 2014 tournament was estimated based of the team of interest. Team of interest was the stronger team (higher seed numbers) in half of the games and the weaker team (lower seed numbers) in the other half of the games.

Differences in ranks of seasonal averages between the teams of all variables found to be significant were placed in the model developed for the first round to verify the accuracy of prediction results for the ordinary least squares regression model. The estimated response $\hat{y}$ then calculated. If $\hat{y}$ is great than 0, a predicted win for the team of interest was coded. If $\hat{y}$ is less than 0, a predicted loss for the team of interest was coded.

Results from the first round of the 2014 tournament were used to validate the first round ordinary least squares regression model and logistic regression model using differences in ranks of seasonal averages. It is noted that the 2014 season was not used in the development of the models.

Table 4.21 gives the results as to how accurately the ordinary least squares regression model for first round of the NCAA 2014 women's volleyball tournament performed.

The first round logistic regression model was validated using the 2014 first round game outcomes and seeing how closely the model agreed. The results are given in Table 4.22.

Table 4.21. Accuracy of ordinary least squares regression model developed by using differences in ranks of seasonal averages when validating first round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	11	3	14
	Loss	9	9	18
	Total	20	12	32
Overall Accuracy				62.5%

Table 4.22. Accuracy of logistic regression model developed by using differences in ranks of seasonal averages when validating first round of 2014

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	11	3	14
	Loss	9	9	18
	Total	20	12	32
Overall Accuracy				62.5%

4.2.6. Validating second round using models developed

Results from the second round of the 2014 tournament were used to validate the second round ordinary least squares regression model and logistic regression model using differences in ranks of seasonal averages. It is noted that the 2014 season was not used in the development of the models.

Table 4.23 gives the results as to how accurately the ordinary least squares regression model for second round of the NCAA 2014 women’s volleyball tournament. Table 4.24 gives equivalent results for the logistic regression model.

Table 4.23. Accuracy of ordinary least squares regression model developed by using differences in ranks of seasonal averages when validating second round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	4	3	7
	Loss	2	7	9
	Total	6	10	16
Overall Accuracy				68.8%

Table 4.24. Accuracy of logistic regression model developed by using differences in ranks of seasonal averages when validating second round of 2014

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	4	3	7
	Loss	2	7	9
	Total	6	10	16
Overall Accuracy				68.8%

4.2.7. Validating third and higher rounds using models developed

Results from the third and higher rounds of the 2014 tournament were used to validate the third and higher rounds ordinary least squares regression model and logistic regression model using differences in ranks of seasonal averages. It is noted that the 2014 season was not used in the development of the models.

Table 4.25 gives the results as to how accurately the ordinary least squares regression model for third and higher rounds of the NCAA 2014 women's volleyball tournament. Table 4.26 gives equivalent results for the logistic regression model.

Table 4.25. Accuracy of ordinary least squares regression model developed by using differences in ranks of seasonal averages when predicting third and higher rounds of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	3	8
	Loss	4	3	7
	Total	9	6	15
Overall Accuracy				53.3%

Table 4.26. Accuracy of logistic regression model developed by using differences in ranks of seasonal averages when validating third and higher rounds of 2014

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	3	8
	Loss	4	3	7
	Total	9	6	15
Overall Accuracy				53.3%

4.2.8. Bracketing the 2015 tournament before tournament begins - Prediction

Results were predicted for every round before the tournament began. Significant differences in ranks of seasonal averages of variables were found for all teams playing each other in the first round and put into first round model. Significant differences of ranks of seasonal averages for each team predicted to play each other in the second round were placed in second round model and winners of this round were predicted. Differences in ranks of seasonal averages of variables found to be significant of teams predicted to play each other in the third round were placed in the third round model and winning teams predicted for this round. This process continued until a winner was selected.

The predicted results were then compared against the actual results for each round of the game for 2015.

4.2.8.1. Examples for each round in 2015 tournament

4.2.8.1.1. Using Ordinary least squares regression model developed by using differences in ranks of seasonal averages

An example for the first round, second round and then third or higher round will be given as to how the ordinary least squares regression model for a particular round in 2015 tournament was used.

4.2.8.1.1.1. Ordinary least squares regression model for first round

The ordinary least squares regression model for first round developed by using differences in ranks of seasonal averages is:

$$\hat{Y} = (-0.01228 * \text{Diff\_Assists}) + (-0.00925 * \text{Diff\_Blocks})$$

Southern California played Cleveland State in the first round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.27.

Table 4.27. Southern California and Cleveland State Statistics

Team	Score	Assists*	Blocks*
Southern California	3	8	37
Cleveland State	1	67	60
Difference	2	-59	-23

* Ranks based on seasonal averages

Using the model above, the game between Southern California and Cleveland State had a predicted point spread of:

$$\hat{y} = (-0.01228 * -59) + (-0.00925 * -23) = 0.94$$

Since ŷ >0 this game was coded as a correctly predicted win for Southern California, who won the game by a score of 3 to 1.

Northern Arizona played San Diego in the first round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.28.

Table 4.28. Northern Arizona and San Diego Statistics

Team	**Score**	**Assists***	**Blocks***
Northern Arizona	0	92	21
San Diego	3	15	41
Difference	-3	77	-20

* Ranks based on seasonal averages

Using the model above, the game between Northern Arizona and San Diego had a predicted point spread of:

$\hat{y} = (-0.01228 * 77) + (-0.00925 * -20) = -0.76$

Since ŷ < 0 this game was coded as a correctly predicted loss for Northern Arizona, who lost the game by a score of 0 to 3.

North Carolina played UNCW in the first round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.29.

Table 4.29. North Carolina and UNCW Statistics

Team	**Score**	**Assists***	**Blocks***
North Carolina	3	72	4
UNCW	0	204	10
Difference	3	-132	-6

* Ranks based on seasonal averages

Using the model above, the game between North Carolina and UNCW had a predicted point spread of:

$\hat{y} = (-0.01228 * -132) + (-0.00925 * -6) = 1.68$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for North Carolina, who won the game by a score of 3 to 0.

Coastal Carolina played Creighton in the first round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.30.

Table 4.30. Coastal Carolina and Creighton Statistics

Team	Score	Assists*	Blocks*
Coastal Carolina	0	81	170
Creighton	3	30	55
Difference	-3	51	115

* Ranks based on seasonal averages

Using the model above, the game between Coastal Carolina and Creighton had a predicted point spread of:

$\hat{y} = (-0.01228 * 51) + (-0.00925 * 115) = -1.69$

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Coastal Carolina, who lost the game by a score of 0 to 3.

Round 1:

Number correct: 25

Number incorrect: 7

Total: 32

4.2.8.1.1.2. Ordinary least squares regression model for second round

The ordinary least squares regression model for second round developed by using differences in ranks of seasonal averages is:

$$\hat{Y} = 0.00834 * \text{Diff\_Digs}$$

BYU played Western Kentucky in the second round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.31.

Table 4.31. BYU and Western Kentucky Statistics

Team	Score	Digs*
BYU	3	188
Western Kentucky	0	160
Difference	3	28

* Ranks based on seasonal averages

Using the model above, the game between BYU and Western Kentucky had a predicted point spread of:

$\hat{y}$ = 0.00834* 28 = 0.23

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for BYU, who won the game by a score of 3 to 0.

Florida played Florida State in the second round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.32.

Table 4.32. Florida and Florida State Statistics

Team	Score	Digs*
Florida	3	236
Florida State	1	234
Difference	2	2

* Ranks based on seasonal averages

Using the model above, the game between Florida and Florida State had a predicted point spread of:

$\hat{y}$ = 0.00834* 2 = 0.02

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Florida, who won the game by a score of 3 to 1.

Round 2:

Number correct: 8

Number incorrect: 8

Total: 16

4.2.8.1.1.3. Ordinary least squares regression model for Third and Higher Rounds

The ordinary least squares regression model for third and higher rounds developed by using differences in ranks of seasonal averages is:

$$\hat{Y} = -0.02617 * \text{Diff\_Match\_W-L\%}$$

Illinois played Minnesota in the third round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.33.

Table 4.33. Illinois and Minnesota Statistics

Team	**Score**	**Match won-lost percentage***
Illinois	0	98
Minnesota	3	11
Difference	-3	87

* Ranks based on seasonal averages

Using the model above, the game between Illinois and Minnesota had a predicted point spread of:

$\hat{y} = -0.02617 * 87 = -2.28$

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Illinois, who lost the game by a score of 0 to 3.

Texas played Florida in the fourth round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.34.

Table 4.34. Texas and Florida Statistics

Team	**Score**	**Match won-lost percentage***
Texas	3	6
Florida	2	28
Difference	1	-22

* Ranks based on seasonal averages

Using the model above, the game between Texas and Florida had a predicted point spread of:

$\hat{y}$ = -0.02617* -22= 0.58

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Texas, who won the game by a score of 3 to 2.

Round 3-6:

Number correct: 6

Number incorrect: 9

Total: 15

It should be noted that only the teams predicted to play each other in the second round were used in the model. The actual teams were not used all the time since predicting was done before the tournament started.

4.2.8.1.2. Using logistic regression model developed by using differences in ranks of seasonal averages

An example will be given as to how the logistic regression model for a particular round was used for each round in 2015 tournament.

4.2.8.1.2.1. Logistic regression model for first round

The logistic regression model for first round developed by using differences in ranks of seasonal averages is:

$$\pi(\text{Diff\_Blocks, Diff\_Kills}) = \frac{e^{-0.0114*\text{Diff\_Blocks}-0.0169*\text{Diff\_Kills}}}{1+e^{-0.0114*\text{Diff\_Blocks}-0.0169*\text{Diff\_Kills}}}$$

Southern California played Cleveland State in the first round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.35.

Table 4.35. Southern California and Cleveland State Statistics

Team	Score	Blocks*	kills*
Southern California	3	37	7
Cleveland State	1	60	56
Difference	2	-23	-49

* Ranks based on seasonal averages

Using the model above, the game between Southern California and Cleveland State had an estimated probability of winning the game of:

$$\pi(-23, -49) = \frac{e^{-0.0114*-23-0.0169*-49}}{1+e^{-0.0114*-23-0.0169*-49}} = 0.75$$

Since π >0.5 this game was coded as a correctly predicted win for Southern California, who won the game by a score of 3 to 1.

Northern Arizona played San Diego in the first round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.36.

Table 4.36. Northern Arizona and San Diego Statistics

Team	Score	Blocks*	kills*
Northern Arizona	0	21	108
San Diego	3	41	9
Difference	-3	-20	99

* Ranks based on seasonal averages

Using the model above, the game between Northern Arizona and San Diego had an estimated probability of winning the game of:

$$\pi(-20, 99) = \frac{e^{-0.0114*-20-0.0169*99}}{1+e^{-0.0114*-20-0.0169*99}} = 0.19$$

Since π < 0.5 this game was coded as a correctly predicted loss for Northern Arizona, who lost the game by a score of 0 to 3.

North Carolina played UNCW in the first round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.37.

Table 4.37. North Carolina and UNCW Statistics

Team	Score	Blocks*	kills*
North Carolina	3	4	58
UNCW	0	10	177
Difference	3	-6	-119

* Ranks based on seasonal averages

Using the model above, the game between North Carolina and UNCW had an estimated probability of winning the game of:

π(-6, -119) = $\frac{e^{-0.0114*-6-0.0169*-119}}{1+e^{-0.0114*-6-0.0169*-119}}$ = 0.89

Since π> 0.5 this game was coded as a correctly predicted win for North Carolina, who won the game by a score of 3 to 0.

Coastal Carolina played Creighton in the first round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.38.

Table 4.38. Coastal Carolina and Creighton Statistics

Team	Score	Blocks*	kills*
Coastal Carolina	0	170	61
Creighton	3	55	43
Difference	-3	115	18

* Ranks based on seasonal averages

Using the model above, the game between Coastal Carolina AND Creighton had an estimated probability of winning the game of:

π(115,18) = $\frac{e^{-0.0114*115-0.0169*18}}{1+e^{-0.0114*115-0.0169*18}}$ = 0.17

Since π< 0.5 this game was coded as a correctly predicted loss for Coastal Carolina, who lost the game by a score of 0 to 3.

Round 1:

Number correct: 25

Number incorrect: 7

Total: 32

4.2.8.1.2.2. Logistic regression model for second round

The logistic regression model for second round developed by using differences in ranks of seasonal averages is:

$$\pi(\text{Diff\_Digs})=\frac{e^{0.00653*\text{Diff\_Digs}}}{1+e^{0.00653*\text{Diff\_Digs}}}$$

BYU played Western Kentucky in the second round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.39.

Table 4.39. BYU and Western Kentucky Statistics

Team	**Score**	**Digs***
BYU	3	188
Western Kentucky	0	160
Difference	3	28

* Ranks based on seasonal averages

Using the model above, the game between BYU and Western Kentucky had an estimated probability of winning the game of:

$$\pi(28)=\frac{e^{0.00653*28}}{1+e^{0.00653*28}}= 0.55$$

Since π>0.5 this game was coded as a correctly predicted win for BYU, who won the game by a score of 3 to 0.

Florida played Florida State in the second round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.40.

Table 4.40. Florida and Florida State Statistics

Team	**Score**	**Digs***
Florida	3	236
Florida State	1	234
Difference	2	2

* Ranks based on seasonal averages

Using the model above, the game between Florida and Florida State had an estimated probability of winning the game of:

$\pi(2) = \frac{e^{0.00653*2}}{1+e^{0.00653*2}} = 0.51$

Since π >0.5 this game was coded as a correctly predicted win for Florida, who won the game by a score of 3 to 1.

Round 2:

Number correct: 6

Number incorrect: 10

Total: 16

4.2.8.1.2.3. Logistic regression model for third and higher rounds

The logistic regression model for third and higher rounds developed by using differences in ranks of seasonal averages is:

$$\pi(\text{Diff\_W-L\%}) = \frac{e^{-0.0295*\text{Diff\_W-L\%}}}{1+e^{-0.0295*\text{Diff\_W-L\%}}}$$

Texas played Florida in the fourth round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.41.

Table 4.41. Texas and Florida Statistics

Team	Score	Match won-lost percentage*
Texas	3	6
Florida	2	28
Difference	1	-22

* Ranks based on seasonal averages

Using the model above, the game between Texas and Florida had an estimated probability of winning the game of:

π(-22)=$\frac{e^{-0.0295*-22}}{1+e^{-0.0295*-22}}$= 0.66

Since π > 0.5 this game was coded as a correctly predicted win for Texas, who won the game by a score of 3 to 2.

Texas played Minnesota in the fifth round of the 2015 tournament. Data on significant differences in ranks of seasonal averages was collected and displayed in Table 4.42.

Table 4.42. Texas and Minnesota Statistics

Team	**Score**	**Match won-lost percentage***
Texas	3	6
Minnesota	1	11
Difference	2	-5

* Ranks based on seasonal averages

Using the model above, the game between Texas and Minnesota had an estimated probability of winning the game of:

π(-5)=$\frac{e^{-0.0295*-5}}{1+e^{-0.0295*-5}}$= 0.54

Since π > 0.5 this game was coded as a correctly predicted win for Texas, who won the game by a score of 3 to 1.

Round 3-6:

Number correct: 6

Number incorrect: 9

Total: 15

It should be noted that only the teams predicted to play each other in the second round were used in the model. The actual teams were not used all the time since predicting was done before the tournament started.

4.2.9. Results for prediction by using models developed by differences in ranks of seasonal averages

Ordinary least square regression models which developed by using differences in ranks of seasonal averages were used to predict each round of NCAA women's volleyball tournament of 2015. A summary of the number of correct and incorrect predictions for each round of the 2015 tournament is given in Table 4.43.

Table 4.43. Prediction results of each round for 2015: (Ordinary least squares regression model developed by using differences in ranks of seasonal averages)

	Correct	**Incorrect**	**Total games**
First round	25	7	32
Second round	8	8	16
Third round	4	4	8
Fourth round	2	2	4
Fifth round	0	2	2
Final round	0	1	1
	Overall Accuracy		61.9%

Logistic regression models which developed by using differences in ranks of seasonal averages were used to predict each round of NCAA women's volleyball tournament of 2015. A summary of the number of correct and incorrect predictions for each round of the 2015 tournament is given in Table 4.44.

Table 4.44. Prediction results of each round for 2015: (Logistic regression model developed by using differences in ranks of seasonal averages)

	Correct	**Incorrect**	**Total games**
First round	25		32
Second round	6		16
Third round	3		8
Fourth round	2		4
Fifth round	1		2
Final round	0		1
	Overall Accuracy		58.7%

It is noted that ordinary least squares regression model works slightly better than logistic regression model when using differences in ranks of seasonal averages to develop models.

4.3. Model developed by using difference of seasonal averages

4.3.1. Develop models by using seasonal averages

All data collected from NCAA.COM [4], seasonal averages were collected before the tournament started. For example, the first game of NCAA 2011 women's volley ball tournament was held on December 1, 2011, the differences of seasonal averages were based on all games through November 27, 2011.

Data was collected for three years of the NCAA women's volleyball tournament. This included 2011, 2012 and 2013 tournaments. Seasonal averages for the variables listed in Table 4.1 (Set A) were collected for all the teams in the 2011 tournament. The variables included: Aces Per Set, Assists Per Set, Blocks Per Set, Digs Per Set, Hitting Percentage, Kills Per Set and Match W-L Percentage. Seasonal averages were also collected on the same variables for all teams playing in the 2012 and 2013 tournament.

4.3.2. Develop models for the first round using seasonal averages

4.3.2.1. Develop ordinary least squares regression models

The response variable for the ordinary least squares regression model was point spread in the order of the team of interest minus the opposing team. Team of interest was the stronger team (higher seed numbers) in half of the games and the weaker team (lower seed numbers) in the other half of the games. A positive point spread indicates a win for the team of interest and a negative value indicates a loss for the team of interest. There were 192 teams playing 96 games in first rounds of the tournaments in 2011, 2012 and 2013. For the first half games of the first round in the three years, the point spread was obtained by using the stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the remainder games of the first round of the three years, the point spread was acquired by using the scores of weaker team (lower seed numbers) minus stronger team (higher seed numbers).

The intercept was excluded when developing the models because the models should give the same results regardless of the ordering of the teams in the model. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences of the seasonal averages for all the variables previously given in Table 4.1 (Set A) between the two teams were considered for entry in the model.

4.3.2.1.1. Development of ordinary least squares regression model for the first round

The ordinary least squares regression model to help predict the winning team for each game in the first round based on using differences of seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-2.04026^* \text{ Diff\_ Aces}) + (28.72233^* \text{ Diff\_Hitting\%})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.45. Table 4.46 gives the steps associated with the stepwise selection technique and Table 4.47 shows the associated R-square values as variables are added to the model. The model with the 2 significant variables explains an estimated 39% of the variation in point spread.

Table 4.45. Point spread model parameter estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Diff_Aces	1	-2.04026	0.55647	-3.67	0.0004	1.00886
Diff_Hitting%_	1	28.72233	3.97361	7.23	<.0001	1.00886

Table 4.46. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Diff_Hitting%		1	0.3080	0.3080	17.4884	42.28	<.0001
2	Diff_Aces		2	0.0866	0.3946	5.5395	13.44	0.0004

Table 4.47. Summary of R-squares value

Root MSE	1.99186	R-Square	0.3946
Dependent Mean	-0.14583	Adj R-Sq	0.3817
Coeff Var	-1365.84572		

4.3.2.2. Develop logistic regression models

The logistic regression model was also fit to the data with the dependent variable recorded as '1' for win and '0' for loss for the team of interest. The model estimates the probability of a win for the team of interest. Team of interest was the stronger team (higher seed numbers) in half of the games and the weaker team (lower seed numbers) in the other half of the games.

No intercept was included during the development of the logistic regression model because the ordering of the teams in the model should not matter. Stepwise selection was used with an α value of 0.1 for both entry and exit when determining the significant variables in developing the logistic regression model. The differences of the seasonal averages for both teams for all previously mentioned variables given in Table 4.1 (Set A) were considered for entry in the model.

4.3.2.2.1. Development of logistic regression model for the first round

A logistic regression model to help predict the winning team for each game in the first round was developed and found to be:

$$\pi(\text{Diff\_Assists, Diff\_Blocks})=\frac{e^{1.3902*\text{Diff\_Assists}+2.5910*\text{Diff\_Blocks}}}{1+e^{1.3902*\text{Diff\_Assists}+2.5910*\text{Diff\_Blocks}}}$$

Where π (Diff_Assists, Diff_Blocks) is the estimated probability that the team of interest will win the game with difference of seasonal averages in assists and difference of seasonal averages in blocks in model.

Table 4.48 shows the steps for the stepwise selection technique and Table 4.49 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 4.50 shows the Hosmer and Lemeshow Test [5] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.6488 indicating that there was no evidence to reject using the logistic regression model.

Table 4.48. Summary of stepwise selection for logistic regression model

Step	**Effect**		**DF**	**Number In**	**Score Chi-Square**	**Wald Chi-Square**	**Pr > ChiSq**	**Variable Label**
	Entered	**Removed**						
1	Diff_Hitting%		1	1	21.8592		<.0001	Diff_Hitting%_
2	Diff_Blocks		1	2	7.4371		0.0064	Diff_Blocks
3	Diff_Assists		1	3	8.1293		0.0044	Diff_Assists
4		Diff_Hitting%	1	2		0.5308	0.4663	Diff_Hitting%_

Table 4.49. Logistic regression model parameter estimates

Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Diff_Assists	1	1.3902	0.3732	13.8739	0.0002
Diff_Blocks	1	2.5910	0.6129	17.8702	<.0001

Table 4.50. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	**DF**	**Pr > ChiSq**
5.9861	8	0.6488

4.3.3. Develop models for the second round using seasonal averages

4.3.3.1. Develop ordinary least squares regression models

There were 96 teams playing 48 games in second rounds of the tournaments in 2011 to 2013. For the first half games of the second round, the point spread was obtained by using the scores of stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using stronger team (higher seed numbers) minus the weaker team (lower seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences of the seasonal averages of the variables listed in Table 4.1 (Set A) between the two teams were considered for entry in the model.

4.3.3.1.1. Development of ordinary least squares regression model for the second round

The ordinary least squares regression model to help predict the winning team for each game in the second round based on using differences of seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = -0.57886 * \text{Diff\_Digs}$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.51. Table 4.52 gives the steps associated with the stepwise selection technique and Table 4.53 shows the associated R-square values as variables are added to the model. The model with the only 1 significant variable explains an estimated 17% of the variation in point spread.

Table 4.51. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Diff_Digs	1	-0.57886	0.18599	-3.11	0.0032	1.00000

Table 4.52. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Diff_Digs		1	0.1709	0.1709	0.3116	9.69	0.0032

Table 4.53. Summary of R-squares value

Root MSE	2.20255	R-Square	0.1709
Dependent Mean	-0.06250	Adj R-Sq	0.1532
Coeff Var	-3524.08603		

4.3.3.2. Develop logistic regression models for the second round

The logistic regression model was also fit for the data with responses recorded as'1' for win and '0' for loss for the team of interest. No intercept was used during the development of the logistic regression model. Stepwise selection was used with an α value of 0.1 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences of the seasonal averages of all previously mentioned variables listed in Table 4.1 (Set A) between the two teams were considered for entry in the model.

4.3.3.2.1. Development of logistic regression model for the second round

A logistic regression model to help predict the winning team for each game in the second round was developed and found to be:

$$\pi(\text{Diff\_Digs})=\frac{e^{-0.5085*\text{Diff\_Digs}}}{1+e^{-0.5085*\text{Diff\_Digs}}}$$

Where π (Diff_Digs) is the estimated probability that the team of interest will win the game with difference of seasonal averages in digs in model.

Table 4.54 shows the steps for the stepwise selection technique and Table 4.55 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 4.56 shows the Hosmer and Lemeshow Test [5] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.8955 indicating that there was no evidence to reject using the logistic regression model.

Table 4.54. Summary of stepwise selection for logistic regression model

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Diff_Digs		1	1	6.4175		0.0113	Diff_Digs

Table 4.55. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Diff_Digs	1	-0.5085	0.2164	5.5214	0.0188

Table 4.56. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
3.5467	8	0.8955

4.3.4. Develop models for the third and higher rounds using seasonal averages

4.3.4.1. Develop ordinary least squares regression models

There were 90 teams playing 45 games in second rounds of the tournaments in 2011 to 2013. For the first 24 games of the second round, the point spread was obtained by using the stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the scores of weaker team (lower seed numbers) minus the stronger team (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences of the seasonal averages of the previously mentioned variables list in Table 4.1 (Set A) between the two teams were considered for entry in the model.

4.3.4.1.1. Development of ordinary least squares regression model for the third and higher rounds

The ordinary least squares regression model to help predict the winning team for each game in the third and higher rounds based on using differences between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = 7.33912 * \text{Diff\_Match\_W-L\%}$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.57. Table 4.58 gives the steps associated with the stepwise selection technique and Table 4.59 shows the associated R-square values as variables are added to the model. The model with this only 1 significant variable explains an estimated 15% of the variation in point spread.

Table 4.57. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Diff_Match_W-L%_	1	7.33912	2.58589	2.84	0.0068	1.00000

Table 4.58. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Diff_Match_W-L%_		1	0.1547	0.1547	0.6923	8.06	0.0068

Table 4.59. Summary of R-squares value

Root MSE	2.18710	R-Square	0.1547
Dependent Mean	0.11111	Adj R-Sq	0.1355
Coeff Var	1968.38887		

4.3.4.2. Develop logistic regression models for the third and higher rounds

The logistic regression model was also fit to the data with responses recorded as‘1’ for win and ‘0’ for loss for the team of interest. No intercept was used during the development of the logistic regression model. Stepwise selection was used with an α value of 0.1 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences of the seasonal averages of all previously mentioned variables listed in Table 4.1 (Set A) between the two teams were considered for entry in the model.

4.3.4.2.1. Development of logistic regression model for the third and higher rounds

A logistic regression model to help predict the winning team for each game in the third and higher rounds was developed and found to be:

$$\pi(\text{Diff\_W-L\%})=\frac{e^{7.2716*\text{Diff\_W-L\%}}}{1+e^{7.2716*\text{Diff\_W-L\%}}}$$

Where π (Diff_W-L%) is the estimated probability that the team of interest will win the game with differences of seasonal averages in win-lose percentage in the model.

Table 4.60 shows the steps for the stepwise selection technique and Table 4.61 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 4.62 shows the Hosmer and Lemeshow Test [5] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.305 indicating that there was no evidence to reject using the logistic regression model.

Table 4.60. Summary of stepwise selection for logistic regression model

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Diff_W-L%		1	1	6.7844		0.0092	Diff_Match_W-L%_

Table 4.61. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Diff_W-L%	1	7.2716	3.0249	5.7789	0.0162

Table 4.62. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
8.3225	7	0.3050

4.3.5. Validating first round using models developed

4.3.5.1. Ordinary least squares regression model

Results from the first round of the 2014 tournament were used to validate the first round ordinary least squares regression model and logistic regression model using differences of seasonal averages. It is noted that the 2014 season was not used in the development of the models.

Table 4.63 gives the results as to how accurately the ordinary least squares regression model for first round of the NCAA 2014 women's volleyball tournament.

Table 4.63. Accuracy of ordinary least squares regression model developed by seasonal averages when validating first round of 2014

Point spread		**Predicted**		
		Win	Loss	Total
Actual	Win	10	4	14
	Loss	7	11	18
	Total	17	15	32
Overall Accuracy				65.6%

The first logistic regression models developed by using seasonal actual average data was used to predict the first round of 2014 season to check the prediction accuracy of the model. It is noted that the 2014 season was not used in the development of the models.

Table 4.64 gives the results as to how accurately the logistic regression model for first round of the NCAA 2014 women's volleyball tournament.

Table 4.64. Accuracy of logistic regression model developed by seasonal averages when validating first round of 2014

Logistic		**Predicted**		
		Win	Loss	Total
Actual	Win	11	4	15
	Loss	6	11	17
	Total	17	15	32
	Overall Accuracy			68.8%

It is noted that the percentage of accuracy for first rounds using the logistic regression models developed by using seasonal average differences works slightly better than the ordinary least squares regression model.

4.3.6. Validating second round using models developed

Results from the second round of the 2014 tournament were used to validate the second round ordinary least squares regression model and logistic regression model using differences of seasonal averages. It is noted that the 2014 season was not used in the development of the models.

Table 4.65 gives the results as to how accurately the ordinary least squares regression model for second round of the NCAA 2014 women's volleyball tournament. Table 4.66 gives equivalent results for the logistic regression model.

Table 4.65. Accuracy of ordinary least squares regression model developed by seasonal averages when validating second round of 2014

Point spread		Predicted		
		Win	Loss	Total
Actual	Win	4	3	7
	Loss	2	7	9
	Total	6	10	16
Overall Accuracy				68.8%

Table 4.66. Accuracy of logistic regression model developed by seasonal averages when validating second round of 2014

Logistic		Predicted		
		Win	Loss	Total
Actual	Win	4	3	7
	Loss	2	7	9
	Total	6	10	16
Overall Accuracy				68.8%

4.3.7. Validating third and higher rounds using models developed

Results from the third and higher rounds of the 2014 tournament were used to validate the third and higher rounds ordinary least squares regression model and logistic regression model using differences of seasonal averages.

Table 4.67 gives the results as to how accurately the ordinary least squares regression model for third and higher rounds of the NCAA 2014 women’s volleyball tournament. Table 4.68 gives equivalent results for the logistic regression model.

Table 4.67. Accuracy of ordinary least squares regression model developed by seasonal averages when validating third and higher rounds of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	3	8
	Loss	4	3	7
	Total	9	6	15
Overall Accuracy				53.3%

Table 4.68. Accuracy of logistic regression model developed by seasonal averages when validating third and higher rounds of 2014

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	3	8
	Loss	4	3	7
	Total	9	6	15
Overall Accuracy				53.33%

4.3.8. Bracketing the 2015 tournament before tournament begins – Prediction

Results were predicted for every round before the 2015 tournament began. Significant differences of seasonal averages of variables were found for all teams playing in the first round and put into first round model. Significant differences of seasonal averages for each team predicted to play each other in the second round were placed in second round model and winners of this round were predicted. Differences of seasonal averages of variables found to be significant of teams predicted to play each other in the third round were placed in the third round model and winning teams predicted for this round. This process continued until a winner was selected.

The predicted results were then compared against the actual results for each round of the game for 2015.

4.3.8.1. Examples for each round in 2015 tournament

An example for the first round, second round and then third or higher round will be given as to how the ordinary least squares regression model for a particular round in 2015 tournament was used.

4.3.8.1.1. Ordinary least squares regression model developed by using seasonal averages

4.3.8.1.1.1. Ordinary least squares regression model for first round

The ordinary least squares regression model for first round developed by using differences of seasonal averages is:

$$\hat{Y} = (-2.04026^{*} \text{ Diff\_ Aces}) + (28.72233^{*} \text{ Diff\_Hitting\%})$$

Southern California played Cleveland State in the first round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.69.

Table 4.69. Southern California and Cleveland State Statistics

Team	Score	Aces*	Hitting percentage*
Southern California	3	1.52	0.292
Cleveland State	1	1.05	0.248
Difference	2	0.47	0.044

* Average per game for season

Using the model above, the game between Southern California and Cleveland State had a predicted point spread of:

$$\hat{y} = (-2.04026^{*}\ 0.47) + (28.72233^{*}\ 0.044) = 0.3$$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Southern California, who won the game by a score of 3 to 1.

Northern Arizona played San Diego in the first round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.70.

Table 4.70. Northern Arizona and San Diego Statistics

Team	Score	Aces*	Hitting percentage*
Northern Arizona	0	1.77	0.264
San Diego	3	1.05	0.22
Difference	-3	0.72	0.044

* Average per game for season

Using the model above, the game between Northern Arizona and San Diego had a predicted point spread of:

$\hat{y} = (-2.04026 * 0.72) + (28.72233 * 0.044) = -0.21$

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Northern Arizona, who lost the game by a score of 0 to 3.

North Carolina played UNCW in the first round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.71.

Table 4.71. North Carolina and UNCW Statistics

Team	Score	Aces*	Hitting percentage*
North Carolina	3	1.12	0.239
UNCW	0	1.22	0.231
Difference	3	-0.1	0.008

* Average per game for season

Using the model above, the game between North Carolina and UNCW had a predicted point spread of:

$\hat{y} = (-2.04026 * -0.1) + (28.72233 * 0.008) = 0.43$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for North Carolina, who won the game by a score of 3 to 0.

Coastal Carolina played Creighton in the first round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.72.

Table 4.72. Coastal Carolina and Creighton Statistics

Team	Score	Aces*	Hitting percentage*
Coastal Carolina	0	1.48	0.289
Creighton	3	1.21	0.249
Difference	-3	0.27	0.04

* Average per game for season

Using the model above, the game between Coastal Carolina and Creighton had a predicted point spread of:

$\hat{y}$ = (-2.04026* 0.27) + (28.72233*0.04) = 0.6

Since $\hat{y} < 0$ this game was coded as an incorrectly predicted win for Coastal Carolina, who actual lost the game by a score of 0 to 3.

Round 1:

Number correct: 25

Number incorrect: 7

Total: 32

4.3.8.1.1.2. Ordinary least squares regression model for second round

The ordinary least squares regression model for second round developed by using differences of seasonal averages is:

$$\hat{Y} = -0.57886 * \text{Diff\_Digs}$$

BYU played Western Kentucky in the second round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.73.

Table 4.73. BYU and Western Kentucky Statistics

Team	Score	Digs*
BYU	3	14.63
Western Kentucky	0	14.96
Difference	3	-0.33

* Average per game for season

Using the model above, the game between BYU and Western Kentucky had a predicted point spread of:

$\hat{y}$ = -0.57886* -0.33 = 0.19

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for BYU, who won the game by a score of 3 to 0.

Florida played Florida State in the second round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.74.

Table 4.74. Florida and Florida State Statistics

Team	Score	Digs*
Florida	3	14.08
Florida State	1	14.12
Difference	2	-0.04

* Average per game for season

Using the model above, the game between Florida and Florida State had a predicted point spread of:

$\hat{y}$ = -0.57886* -0.04 = 0.02

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Florida, who won the game by a score of 3 to 1.

Round 2:

Number correct: 8

Number incorrect: 8

Total: 16

4.3.8.1.1.3. Ordinary least squares regression model for third and higher rounds

The ordinary least squares regression model for third and higher rounds developed by using differences of seasonal averages is:

$$\hat{Y} = 7.33912 * \text{Diff\_Match\_W-L\%}$$

Texas played Florida in the fourth round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.75.

Table 4.75. Texas and Florida Statistics

Team	Score	Match won-lost percentage*
Texas	3	0.929
Florida	2	0.793
Difference	1	0.136

* Average per game for season

Using the model above, the game between Texas and Florida had a predicted point spread of:

$\hat{y} = 7.33912 * 0.136 = 0.99$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Texas, who won the game by a score of 3 to 2.

Texas played Minnesota in the fifth round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.76.

Table 4.76. Texas and Minnesota Statistics

Team	Score	Match won-lost percentage*
Texas	3	0.929
Minnesota	1	0.867
Difference	2	0.062

* Average per game for season

Using the model above, the game between Texas and Minnesota had a predicted point spread of:

$\hat{y}$ = 7.33912* 0.062 = 0.46

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Texas, who won the game by a score of 3 to 1.

Round 3-6:

Number correct: 7

Number incorrect: 8

Total: 15

It should be noted that only the teams predicted to play each other in the second round were used in the model. The actual teams were not used all the time since predicting was done before the tournament started.

4.3.8.1.2. Using logistic regression model developed by seasonal averages

An example will be given as to how the logistic regression model for a particular round was used for each round in 2015 tournament.

4.3.8.1.2.1. Logistic regression model for first round

The logistic regression model for first round developed by using differences of seasonal averages is:

$$\pi(\text{Diff\_Assists, Diff\_Blocks}) = \frac{e^{1.3902*\text{Diff\_Assists}+2.5910*\text{Diff\_Blocks}}}{1+e^{1.3902*\text{Diff\_Assists}+2.5910*\text{Diff\_Blocks}}}$$

Southern California played Cleveland State in the first round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.77.

Table 4.77. Southern California and Cleveland State Statistics

Team	Score	Assists*	Blocks*
Southern California	3	13.75	2.53
Cleveland State	1	12.76	2.38
Difference	2	0.99	0.15

* Average per game for season

Using the model above, the game between Southern California and Cleveland State had an estimated probability of winning the game of:

$$\pi(0.99, 0.15) = \frac{e^{1.3902*0.99+2.5910*0.15}}{1+e^{1.3902*0.99+2.5910*0.15}} = 0.85$$

Since π >0.5 this game was coded as a correctly predicted win for Southern California, who won the game by a score of 3 to 1.

Northern Arizona played San Diego in the first round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.78.

Table 4.78. Northern Arizona and San Diego Statistics

Team	Score	Assists*	Blocks*
Northern Arizona	0	12.42	2.79
San Diego	3	13.48	2.51
Difference	-3	-1.06	0.28

* Average per game for season

Using the model above, the game between Northern Arizona and San Diego had an estimated probability of winning the game of:

$$\pi(-1.06, 0.28) = \frac{e^{1.3902*-1.06+2.5910*0.28}}{1+e^{1.3902*-1.06+2.5910*0.28}} = 0.32$$

Since π < 0.5 this game was coded as a correctly predicted loss for Northern Arizona, who lost the game by a score of 0 to 3.

North Carolina played UNCW in the first round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.79.

Table 4.79. North Carolina and UNCW Statistics

Team	Score	Assists*	Blocks*
North Carolina	3	12.68	3.07
UNCW	0	11.41	2.95
Difference	3	1.27	0.12

* Average per game for season

Using the model above, the game between North Carolina and UNCW had an estimated probability of winning the game of:

$$\pi(1.27, 0.12) = \frac{e^{1.3902*1.27+2.5910*0.12}}{1+e^{1.3902*1.27+2.5910*0.12}} = 0.89$$

Since $\pi > 0.5$ this game was coded as a correctly predicted win for North Carolina, who won the game by a score of 3 to 0.

Coastal Carolina played Creighton in the first round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.80.

Table 4.80. Coastal Carolina and Creighton Statistics

Team	Score	Assists*	Blocks*
Coastal Carolina	0	12.56	2.03
Creighton	3	13.18	2.39
Difference	-3	-0.62	-0.36

* Average per game for season

Using the model above, the game between Coastal Carolina and Creighton had an estimated probability of winning the game of:

$$\pi(-0.62, -0.36) = \frac{e^{1.3902*-0.62+2.5910*-0.36}}{1+e^{1.3902*-0.62+2.5910*-0.36}} = 0.14$$

Since $\pi < 0.5$ this game was coded as a correctly predicted loss for Coastal Carolina, who lost the game by a score of 0 to 3.

Round 1:

Number correct: 26

Number incorrect: 6

Total: 32

4.3.8.1.2.2. Logistic regression model for second round

The logistic regression model for second round developed by using differences of seasonal averages is:

$$\pi(\text{Diff\_Digs}) = \frac{e^{-0.5085*\text{Diff\_Digs}}}{1+e^{-0.5085*\text{Diff\_Digs}}}$$

BYU played Western Kentucky in the second round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.81.

Table 4.81. BYU and Western Kentucky Statistics

Team	Score	Digs*
BYU	3	14.63
Western Kentucky	0	14.96
Difference	3	-0.33

* Average per game for season

Using the model above, the game between BYU and Western Kentucky had an estimated probability of winning the game of:

$$\pi(-0.33) = \frac{e^{-0.5085*-0.33}}{1+e^{-0.5085*-0.33}} = 0.54$$

Since π>0.5 this game was coded as a correctly predicted win for BYU, who won the game by a score of 3 to 0.

Florida played Florida State in the second round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.82.

Table 4.82. Florida and Florida State Statistics

Team	Score	Digs*
Florida	3	14.08
Florida State	1	14.12
Difference	2	-0.04

* Average per game for season

Using the model above, the game between Florida and Florida State had an estimated probability of winning the game of:

π(-0.04) = $\frac{e^{-0.5085*-0.04}}{1+e^{-0.5085*-0.04}}$ = 0.51

Since π >0.5 this game was coded as a correctly predicted win for Florida, who won the game by a score of 3 to 1.

Round 2:

Number correct: 9

Number incorrect: 7

Total: 16

4.3.8.1.2.3. Logistic regression model for third and higher rounds

The logistic regression model for third and higher rounds developed by using differences of seasonal averages is:

π(Diff_W-L%) = $\frac{e^{7.2716*\text{Diff\_W-L\%}}}{1+e^{7.2716*\text{Diff\_W-L\%}}}$

BYU played Nebraska in the third round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.83.

Table 4.83. BYU and Nebraska Statistics

Team	Score	Match won-lost percentage*
BYU	0	0.897
Nebraska	3	0.867
Difference	-3	0.03

* Average per game for season

Using the model above, the game between BYU and Nebraska had an estimated probability of winning the game of:

$$\pi(0.03) = \frac{e^{7.2716*0.03}}{1+e^{7.2716*0.03}} = 0.55$$

Since $\pi > 0.5$ this game was coded as an incorrectly predicted win for BYU, who lost the game by a score of 0 to 3.

Texas played Minnesota in the fifth round of the 2015 tournament. Data on significant differences of seasonal averages was collected and displayed in Table 4.84.

Table 4.84. Texas and Minnesota Statistics

Team	Score	Match won-lost percentage*
Texas	3	0.929
Minnesota	1	0.867
Difference	2	0.062

* Average per game for season

Using the model above, the game between Texas and Minnesota had an estimated probability of winning the game of:

$$\pi(0.062)=\frac{e^{7.2716*0.062}}{1+e^{7.2716*0.062}} = 0.61$$

Since $\pi > 0.5$ this game was coded as a correctly predicted win for Texas, who won the game by a score of 3 to 1.

Round 3-6:

Number correct: 6

Number incorrect: 9

Total: 15

4.3.9. Results for prediction by using models developed by difference of seasonal averages

In 2015, a continuous process was used in verifying the models instead of doing round by round predictions as in 2014. In other words, a complete bracket was filled out in 2015 before any game was played.

The ordinary least squares regression model for the first round developed by using actual seasonal averages was used to predict the teams who go to next round. Once the teams in the second round were predicted, the second-round models were used to predict the winners of the second round. This process was continued for the third and higher rounds until the predicted final winner of the game was determined.

Accuracy of ordinary least squares regression model results was given in Table 4.85 and results of logistic regression models was given in Table 4.86.

Table 4.85. Prediction results of each round for 2015: (Ordinary least squares regression model developed by seasonal averages)

	Correct	**Incorrect**	**Total games**
First round	25	7	32
Second round	8	8	16
Third round	4	4	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	0	1	1
	Overall Accuracy		63.5%

Table 4.86. Prediction results of each round for 2015: (Logistic regression model developed by seasonal averages)

	Correct	**Incorrect**	**Total games**
First round	26	6	32
Second round	9	7	16
Third round	3	5	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	0	1	1
	Overall Accuracy		65.1%

It is noted logistic regression model worked slightly better than ordinary least squares regression model when using seasonal averages to develop models on this data set.

4.4. Model developed by using difference of in-game statistics

4.4.1. Develop models by using in-game statistics

Data was collected for NCAA women's volleyball tournament of 2015. In-game statistics were collected for 37 games of 63 games of the 2015 tournament on the variables listed in Table 4.2 (Set B). The variables included: Attack K, Attack E, Attack Percentage, SERVE SA, SRV RE, Digs and Blocks.

4.4.1.1. Develop ordinary least squares regression model

The response variable for the ordinary least squares regression model was point spread in the order of the team of interest minus the opposing team. Team of interest was the stronger team (higher seed numbers) in half of the games and the weaker team (lower seed numbers) in the other half of the games. A positive point spread indicates a win for the team of interest and a negative value indicates a loss for the team of interest. There were 64 teams playing 63 games in the tournaments of 2015. However, only 37 games have the in-game statistics data. For the first 18 games of 2015 years, the point spread was obtained by using stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the other 17 games, the point spread was acquired by using weaker team (lower seed numbers) minus the stronger team (higher seed numbers).

The intercept was excluded when developing the models because the models should give the same results regardless of the ordering of the teams in the model. Stepwise selection was used with an α value of 0.1 for both entry and exit to develop the models. The differences of the in-game statistics for all the variables previously given in Table 4.2 (Set B) between the two teams were considered for entry in the model.

The ordinary least squares regression model to help explain the variation in point spread for each game in first round through final round based on using differences between in-game statistics of the significant variables was developed and found to be:

$\hat{Y}$ = (0.04538* Diff_AttackK) + (8.12106 * Diff_Attack%)+ (0.21009* Diff_ ServeSA)

The following statistics have positive coefficients associated with them which is to be expected: Difference in Attack Kills, Diff in Attack Percentage and Difference in Serve SA. It is noted that if the team increases Attack Percentage by 1% more than the other team, on average the team will get approximately 0.08 more points. Each additional Attack Kill over the other team is worth approximately 0.05 points.

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.87. Table 4.88 gives the steps associated with the stepwise selection technique and Table 4.89 shows the associated R-square values as variables are added to the model. The model with the 3 significant variables explains an estimated 82% of the variation in point spread.

Table 4.87. Point spread model parameter estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Attack_K	1	0.04538	0.02365	1.92	0.0634	2.40263
Attack_PCT	1	8.12106	1.71816	4.73	<.0001	2.43370
SERVE_SA	1	0.21009	0.04758	4.42	<.0001	1.19471

Table 4.88. Summary of stepwise selection for point spread model

Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Attack_PCT		1	0.6883	0.6883	24.1946	79.48	<.0001
2	SERVE_SA		2	0.1154	0.8036	4.2859	20.57	<.0001
3	Attack_K		3	0.0192	0.8228	2.6413	3.68	0.0634

Table 4.89. Summary of R-squares value

Root MSE	1.02086	R-Square	0.8228
Dependent Mean	-0.05405	Adj R-Sq	0.8072
Coeff Var	-1888.59651		

4.4.1.2. Develop logistic regression model using in-game statistics

The logistic regression model was also fit for the data with responses recorded as'1' for win and '0' for loss for the team of interest. No intercept was used during the development of the logistic regression model. Stepwise selection was used with an α value of 0.1 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences of the in-game statistics of all previously mentioned variables listed in Table 4.2 (Set B) between the two teams were considered for entry in the model.

A logistic regression model to estimate the probability of the team of interest winning based on in-game statistics for each game in round 1 through final round was developed and found to be:

$$\pi(\text{Diff\_AttackPCT, Diff\_ServeSA})=\frac{e^{50.2967*\text{Diff\_AttackPCT}+0.671*\text{Diff\_ServeSA}}}{1+e^{50.2967*\text{Diff\_AttackPCT}+0.671*\text{Diff\_ServeSA}}}$$

Where π (Diff_AttackPCT, Diff_ServeSA) is the estimated probability that the team of interest will win the game with difference of in-game statistics in attack percentage and difference of in-game statistics in serve SA in model.

Table 4.90 shows the steps for the stepwise selection technique and Table 4.91 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 4.92 shows the Hosmer and Lemeshow Test [5] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.907 indicating that there was no evidence to reject using the logistic regression model.

Table 4.90. Summary of stepwise selection for logistic regression model

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Diff_AttackPCT		1	1	20.0456		<.0001	Attack PCT
2	Diff_ServeSA		1	2	6.8386		0.0089	SERVE SA
3	Diff_Blocks		1	3	4.5635		0.0327	Block BS+BA
4		Diff_Blocks	1	2		1.0029	0.3166	Block BS+BA

Table 4.91. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Diff_AttackPCT	1	50.2967	26.8148	3.5183	0.0607
Diff_ServeSA	1	0.6710	0.3534	3.6058	0.0576

Table 4.92. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
2.7513	7	0.9070

4.4.2. Validating first round using models developed

4.4.2.1. Verification of the models developed by using in-game statistics

Using the ordinary least squares regression model developed with in-game statistics, the point spread of each of the 63 games in the 2014 tournament was estimated.

To verify the accuracy of prediction results for the ordinary least squares regression model, values of the in-game statistics were placed in the model for each game. The model result was calculated and compared to the actual result for each game. The estimated response $\hat{y}$ was observed. If $\hat{y}$ was greater than 0, a predicted win for the team of interest was coded. If $\hat{y}$ was less than 0, a predicted loss for the team of interest was coded.

To verify the accuracy of prediction results for the logistic regression model for the first round, a similar process was conducted. For each round of the game, values for the significant in-game statistics were collected and the difference were taken and placed into the logistic regression model to find an estimated probability, $\pi\ x_i$. If π xi was greater than 0.5, a predicted win was coded for the team of interest. If π xi was less than 0.5, a predicted loss was coded for the team of interest.

Results from the first to final rounds of the 2014 tournament were used to validate the ordinary least squares regression model and logistic regression model using differences in in-game statistics. The validation results for the first round using the ordinary least squares regression model and the logistic regression model are given in Table 4.93 and Table 4.94, respectively.

Table 4.93. Accuracy of ordinary least squares regression model developed by in-game statistics when validating first round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	13	2	15
	Loss	2	15	17
	Total	15	17	32
Overall Accuracy				87.5%

Table 4.94. Accuracy of logistic regression model developed by in-game statistics when validating first round of 2014

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	13	2	15
	Loss	2	15	17
	Total	15	17	32
Overall Accuracy				87.5%

It is noted that the percentage of accuracy for first round using the ordinary least squares models and using the logistic regression models which developed by in-game statistics are the same.

4.4.3. Validating second round using models developed

The validation results for second round using ordinary least squares regression model and logistic regression model are given in Table 4.95 and Table 4.96, respectively.

Table 4.95. Accuracy of ordinary least squares regression model developed by in-game statistics when validating second round of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	6	0	6
	Loss	1	9	10
	Total	7	9	16
Overall Accuracy				93.75%

Table 4.96. Accuracy of logistic regression model developed by in-game statistics when validating second round of 2014

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	6	0	6
	Loss	1	9	10
	Total	7	9	16
Overall Accuracy				93.75%

It is noted that the percentage of accuracy for second round using the ordinary least squares models and using the logistic regression models which developed by in-game statistics are the same.

4.4.4. Validating third and higher rounds using models developed

The validation results for third and higher rounds using ordinary least squares regression model and logistic regression model are given in Table 4.97 and Table 4.98, respectively.

Table 4.97. Accuracy of ordinary least squares regression model developed by in-game statistics when validating third and higher rounds of 2014

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	0	5
	Loss	1	9	10
	Total	6	9	15
Overall Accuracy				93.33%

Table 4.98. Accuracy of logistic regression model developed by in-game statistics when validating third and higher rounds of 2014

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	5	0	5
	Loss	0	10	10
	Total	5	10	15
Overall Accuracy				100%

It is noted that the percentage of accuracy for third and higher rounds using the logistic regression models is slightly higher than the percentage of accuracy for ordinary least squares regression models that developed by in-game statistics for this data set.

4.4.5. Bracketing the 2016 tournament before tournament begins – Predicting

Since the in-game statistics will not be available before the tournament begins, differences of seasonal averages of the current year for both teams playing were collected and put into the in-game model to predict the winner of a volleyball game in the 2016 tournament.

Results were predicted for each round by using the ordinary least squares regression model developed using in-game statistics before the 2016 tournament begin by replacing differences between the in-game statistics with differences in seasonal averages.

Differences of seasonal averages of significant variables were found for all teams playing in the first round and put into first round model. Differences of seasonal averages for each team predicted to play each other in the second round were then placed in the model and winners of this round were predicted. Differences of seasonal averages of variables found to be significant of teams predicted to play each other in the third round were placed in the model and winning teams predicted for this round. This process continued until a winner was selected.

The predicted results were then compared against the actual results for each round of the game for 2016.

4.4.5.1. Examples for each round in 2016 tournament

An example for the first round, second round and then third or higher round will be given as to how the ordinary least squares regression model for a particular round in 2016 tournament was used.

4.4.5.1.1. Using ordinary least squares regression model developed by in-game statistics

4.4.5.1.1.1. Ordinary least squares regression model for the whole tournament

The ordinary least squares regression model for first to final round developed by using differences in in-game statistics is:

$\hat{Y}$ = (0.04538* Diff_AttackK) + (8.12106 * Diff_Attack%) + (0.21009* Diff_ ServeSA)

Nebraska played New Hampshire in the first round of the 2016 tournament. Data on differences of seasonal averages for significant variables were collected and displayed in Table 4.99.

Table 4.99. Nebraska and New Hampshire Statistics

Team	**Score**	**Attack_K***	**Attack_Percentage***	**Serve_SA***
Nebraska	3	14.52	0.274	1.09
New Hampshire	0	11.82	0.198	1.73
Difference	3	2.7	0.076	-0.64

* Average per game for season

Using the model above, the game between Nebraska and New Hampshire had a predicted point spread of:

$\hat{y}$ = (0.04538* 2.7) + (8.12106*0.076) + (0.21009* -0.64) = 0.61

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Nebraska, who won the game by a score of 3 to 0.

Kentucky played Colorado State in the first round of the 2016 tournament. Data on differences of seasonal averages for significant variables were collected and displayed in Table 4.100.

Table 4.100. Kentucky and Colorado State Statistics

Team	**Score**	**Attack_K***	**Attack_Percentage***	**Serve_SA***
Kentucky	3	13.65	0.228	1.31
Colorado State	1	14.05	0.273	1.13
Difference	2	-0.4	-0.045	0.18

* Average per game for season

Using the model above, the game between Kentucky and Colorado State had a predicted point spread of:

$\hat{y}$ = (0.04538* -0.4) + (8.12106 *-0.045) + (0.21009* 0.18) = -0.35

Since $\hat{y}$ < 0 this game was coded as an incorrectly predicted loss for Kentucky, who actually won the game by a score of 3 to 1.

Kansas played Samford in the first round of the 2016 tournament. Data on differences of seasonal averages for significant variables were collected and displayed in Table 4.101.

Table 4.101. Kansas and Samford Statistics

Team	Score	Attack_K*	Attack_Percentage*	Serve_SA*
Kansas	3	15.1	0.299	1.32
Samford	0	13.1	0.229	1.47
Difference	3	2	0.07	-0.15

* Average per game for season

Using the model above, the game between Kansas and Samford had a predicted point spread of:

$\hat{y}$ = (0.04538* 2) + (8.12106 *-0.07) + (0.21009* -0.15) = 0.63

Since $\hat{y}$ > 0 this game was coded as a correctly predicted win for Kansas, who won the game by a score of 3 to 0.

UNI played Creighton in the first round of the 2016 tournament. Data on differences of seasonal averages for significant variables were collected and displayed in Table 4.102.

Table 4.102. UNI and Creighton Statistics

Team	Score	Attack_K*	Attack_Percentage*	Serve_SA*
UNI	2	13.36	0.187	1.07
Creighton	3	14.01	0.248	1.16
Difference	-1	-0.65	-0.061	-0.09

* Average per game for season

Using the model above, the game between UNI and Creighton had a predicted point spread of:

$\hat{y} = (0.04538 * -0.65) + (8.12106 * -0.061) + (0.21009 * -0.09) = -0.54$

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for UNI, who lost the game by a score of 2 to 3.

Round 1-6:

Number correct: 30

Number incorrect: 33

Total: 63

It should be noted that only the teams predicted to play each other in the second round were used in the model. The actual teams were not used all the time since predicting was done before the tournament started.

4.4.6. Results for Prediction by using models developed by in-game statistics

In 2016, a continuous process was used in verifying the models instead of doing round by round validations as in 2015. In other words, a complete bracket was filled out in 2016 by using the ordinary least squares regression model and the logistic regression model before any game was played.

The ordinary least squares regression model that was developed by using in-game statistics was used to predict the team in the first round who go to next round. Once the teams in the second round were predicted, the same model was used to predict the winners of the second round. This process was continued for the third and higher rounds until the predicted final winner of the game was determined.

A similar process was conducted for logistic regression model. However, the accuracy is even lower than ordinary least squares regression model, so only ordinary least squares regression model was used to fill out the bracket of 2016 season.

The prediction results for each round of 2016 tournament using ordinary least squares regression model is given in Table 4.103.

Table 4.103. Prediction results of each round for 2015: (Ordinary least squares regression model developed by in-game statistics)

	Correct	**Incorrect**	**Total games**
First round	21	11	32
Second round	6	10	16
Third round	3	5	8
Fourth round	0	4	4
Fifth round	0	2	2
Final round	0	1	1
Overall Accuracy			47.62%

Accuracy of ordinary least squares regression model results was given in Table 4.103 and it is noted the accuracy is only 47.62%. It should be noted that only the teams predicted to play each other in the second round were used in the model. The actual teams were not used all the time since predicting was done before the tournament started.

4.5. Conclusion

4.5.1. Validation - Models developed by using seasonal averages

To verify the accuracy of prediction results for the ordinary least squares regression model, differences of the seasonal averages for both teams for all previously mentioned significant variables were placed in the models developed for the whole tournament. The ordinary least squares regression model developed by using differences in ranks of seasonal averages and seasonal averages for the first round had approximately a 62.5% and a 65.6% chance of correctly predicting the results, respectively. The logistic regression model developed by using differences in ranks of seasonal averages and seasonal averages for the first round had approximately a 62.5% and 68.8% chance of correctly predicting the results, respectively. The ordinary least squares regression model and the logistic regression model developed by using differences in ranks of seasonal averages and seasonal averages for the second round both had approximately a 68.8% chance of correctly predicting the results. The logistic regression model developed by using differences in ranks of seasonal averages and seasonal averages for the second round both had approximately a 68.8% chance of correctly predicting the results. The ordinary least squares regression model and the logistic regression model developed by using differences in ranks of seasonal averages and seasonal averages for the third and higher rounds both had approximately a 53.3% chance of correctly predicting the results. The logistic regression model developed by using differences in ranks of seasonal averages and seasonal averages for the third and higher rounds had approximately a 53.3% chance of correctly predicting the results.

4.5.2. Prediction - Models developed by using seasonal averages

In 2015, a continuous process was used to predict the winning team in each round before the tournament started instead of doing round by round predictions as in 2014. Namely, a complete bracket was filled out in 2015 before any game was played. When the differences of the seasonal averages for both teams for all the significant variables were considered for entry in the ordinary least squares models which developed by using differences in ranks of seasonal averages and differences of seasonal averages, the models had approximately a 61.9% and 63.5% chance of correctly predicting the winner of a volleyball game, respectively. The logistic regression model developed by using differences in ranks of seasonal averages and seasonal averages had approximately a 58.7% and 65.1% chance of correctly predicting the women's volleyball game, respectively.

4.5.3. Validation - Models developed by using in-game statistics

To verify the accuracy of prediction results for the ordinary least squares regression model developed by using in-game statistics, differences of the in-game statistics for both teams for all previously mentioned significant variables were placed in the model developed for the whole tournament. The ordinary least squares regression model and the logistic regression model for the first round both had approximately a 87.5% chance of correctly predicting the results. The ordinary least squares regression model and the logistic regression model for the second round both had approximately a 93.8% chance of correctly predicting the results. The ordinary least squares regression model and the logistic regression model for the third and higher rounds had approximately a 93.33% and a 100% chance of correctly predicting the results, respectively.

It is noted the validation accuracy is high, both ordinary least squares regression model and logistic regression model work great on explain the variables when the model is developed by using in-game statistics.

4.5.4. Prediction - Models developed by using in-game statistics

When the differences of the seasonal averages for both teams for all significant variables were considered for entry in the ordinary least squares regression model developed by using differences of in-game statistics, the model had approximately a 47.6% chance of correctly predicting the winner of a volleyball game.

It is noted that the prediction were done and brackets filled out before the tournament began. The accuracy is lower because teams predicted to play in the second round or higher round might not have actually made it to those rounds.

4.5.5. Overall comparisons

Both the ordinary least squares regression and logistic regression model developed by using in-game statistics work well when the in-game statistics are known.

When predicting results for future tournaments without in-game statistics given, the models developed by using seasonal averages is better than the models developed by in-game statistics. This is not surprising since the model developed using seasonal averages is developed with seasonal averages in mind. The model developed using in-game statistics is not, and then replacing in-game statistics with seasonal averages.

It is noted using difference of seasonal averages is better than using differences in ranks of averages for both ordinary least squares regression model and logistic regression model.

Overall, the logistic regression model developed by using seasonal averages with an overall accuracy 65% works slightly better than the ordinary least squares regression model when predicting the winner of 2015 NCAA women’s volleyball tournament.

4.6. References

[1] NCAA Division I Women's Volleyball Championship. Retrieved October 10, 2017, from https://en.wikipedia.org/wiki/NCAA_Division_I_Women's_Volleyball_Championship

[2] Road to the Championship. Retrieved October 10, 2017, from http://www.ncaa.com/championships/volleyball-women/d1/road-to-the-championship

[3] 2017 NCAA Official Volleyball Statistics Rules. Retrieved October 10, 2017, from http://fs.ncaa.org/Docs/stats/Stats_Manuals/VB/2017.pdf

[4] Ranking Summary. Retrieved October 10, 2016, from http://web1.ncaa.org/stats/StatsSrv/ranksummary

[5] Hosmer, David W.; Lemeshow, Stanley (2013). Applied Logistic Regression.
New York: Wiley. ISBN 978-0-470-58247-3

CHAPTER 5. BRACKETING NCAA WOMEN’S SOCCER TOURNAMENT

5.1. Introduction

5.1.1. The history of NCAA women’s soccer tournament

The NCAA division I women’s soccer tournament is the annual championship in women’s soccer from teams in division I contested by the National Collegiate Athletic Association(NCAA) each winter. It is also known as the Women’s College Cup. There were only 12 team competing for the single division women’s soccer Championship tournament which held in 1982. The tournament became the Division I Championship in 1986. The tournament expanded gradually, and is currently at 64 teams (NCAA - Soccer [1]).

5.1.2. The playing rule and structure

All division I women’s soccer programs were eligible to qualify for the tournament. Twenty-eight teams receive automatic bids by winning their conference tournaments, 3 teams receive automatic bids by claiming the conference regular season crown and the remainder of the teams earn at-large bids based on their regular season records (Road to the Championship [2]).

There are 64 teams playing 32 games to compete in a single elimination tournament for the first round of the NCAA division I women’s soccer tournament championship. The 32 advancing teams then compete against each other in single-elimination second round competition. The winning teams advance to the regional round. For the regional round, there will be 16 teams competing in a single-elimination regional semifinal competition. The advancing teams then compete against each other in single-elimination regional final. The winning team in each of the four regions advanced to the semifinal. The winner of each game in the semifinal advances to the final round and plays for the championship (Road to the Championship [2]). Figure 3 shows the 2015-2016 NCAA women’s soccer tournament bracket.

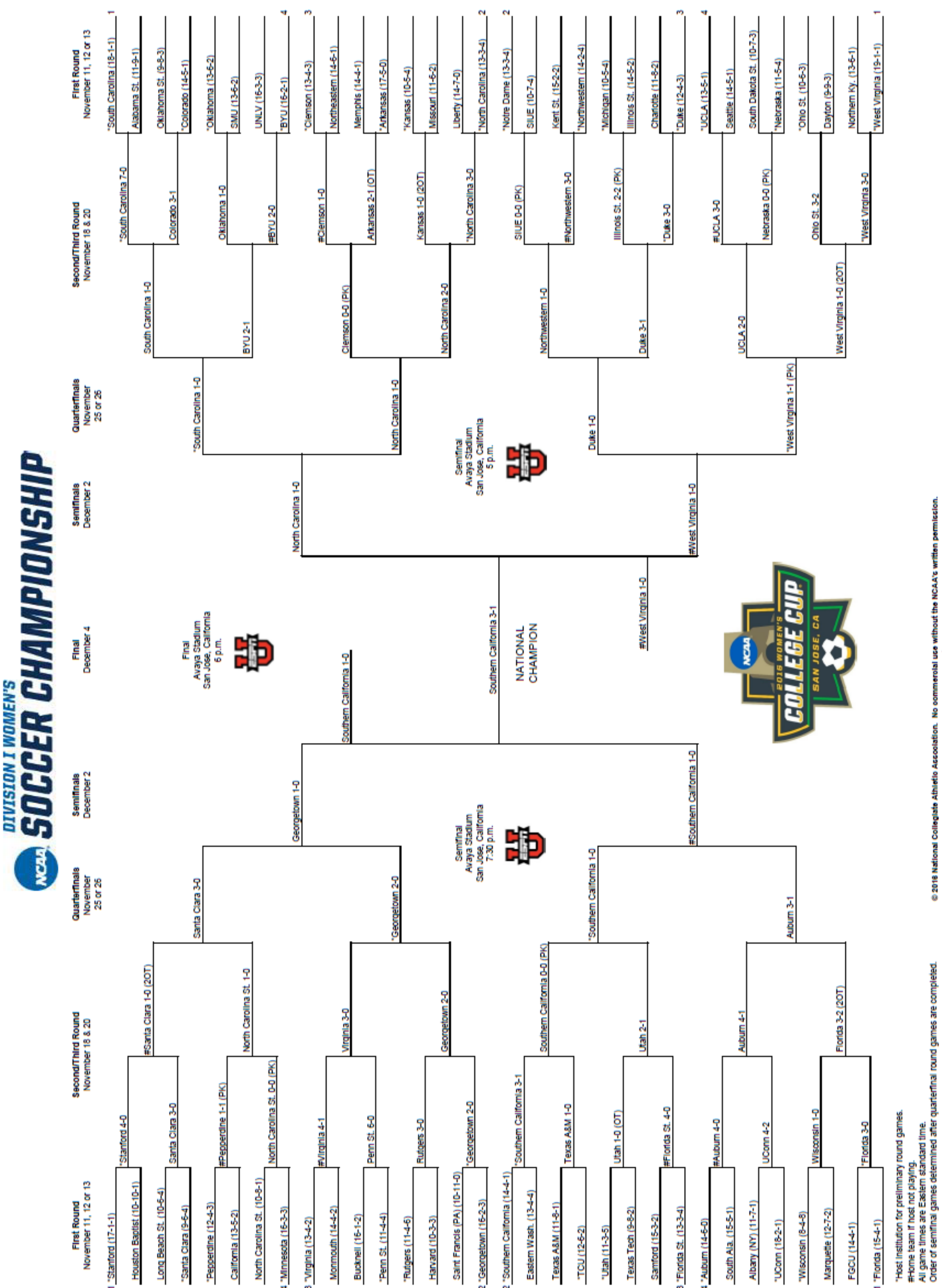

Figure 3. The NCAA women's soccer tournament bracket for the 2015 – 2016 season. (This bracket is downloaded from: http://www.ncaa.com/interactive-bracket/soccer-women/d1)

5.1.3. The research objectives for this study

The research objectives for this study are as follows:

1) Develop ordinary least squares regression models for Round 1, Round 2 and Rounds 3-6 with point spread being the dependent variable by using seasonal average, to predict winners of soccer games in each of those rounds for the NCAA women's soccer tournament; and

2) Develop logistic regression models for Round 1, Round 2 and Rounds 3-6 that estimate the probability of a team winning the game by using seasonal averages, to predict winners of soccer games in each of those rounds for the NCAA women's soccer tournament; and

3) Develop one ordinary least squares regression model by using in-game statistics, to explain the variation of the point spread of a women's soccer game and then use this model to predict the winners of the soccer games for the NCAA women's soccer tournament by replacing the significant in-game statistics with seasonal averages; and

4) Develop one logistic regression model that estimate the probability of a team winning the game by using in-game statistics, and then use this model to predict winners by replacing significant in-game statistics with seasonal averages.

In order to accomplish objectives 1 and 2, data was collected for three years of the NCAA women's soccer tournament. This included the 2013, 2014 and 2015 tournaments. Differences of seasonal averages were collected for all the teams in the 2013 tournament on the variables listed in Table 5.1 (Set A): Scoring Offense, Goals-Against Average, Shutout Percentage, Won-Lost-Tied Percentage, Save Percentage, Saves Per game, Assists Per Game and Points Per Game. Seasonal averages were also collected on the same variables for all teams playing in the 2014 and 2015 tournaments. The developed models are given in Section 5.2.

Table 5.1. Set A - Variables in consideration for seasonal average

Variables in consideration	Definitions
Scoring Offense	SO = $\frac{Total\ Goals}{Total\ Team\ Games\ Played}$ [3]
Goals-Against Average	GAA = $\frac{Goals\ Allowed \times 90}{Total\ Minutes\ Played}$ [4]
Shutout Percentage	Shutout %= $\frac{Total\ Shutouts}{Total\ Team\ Games\ Played}$ [3]
Won-Lost-Tied Percentage	WLT %= $\frac{Wins+(\frac{1}{2}\ of\ ties)}{Total\ Team\ Games\ Played}$ [3]
Save Percentage	Save %=$\frac{Saves}{Saves+Goals\ Allowed}$ [4]
Saves Per Game	SPG = $\frac{Total\ Saves}{Total\ Games\ Played}$ [3]
Assists Per Game	APG = $\frac{Total\ Assists}{Total\ Games\ Played}$ [4]
Points Per Game	PPG = $\frac{Total\ Points}{Total\ Games\ Played}$ [4]

For research objectives 3 and 4, data was collected for NCAA women's soccer tournament of 2016. In-game statistics were collected for all the games in the 2016 tournament on the variables listed in Table 5.2 (Set B). The variables included: Shutout, SOG, Assists, Fouls, Goalie Saves and Offside. The developed models are given in Section 5.3.

Table 5.2. Set B - Variables in consideration for in-game statistics

Variables in consideration	Definitions
Shutout	A shut out is earned when the opposite team fails to score a single goal during a game. [5]
SOG	Any time a player makes an attempt to take a shot that does or would enter the goal is considered a (SOG). This includes shots that bounce off the goals, shots stopped by a defender, or shots saved by a goalkeeper. [5]

Table 5.2. Set B - Variables in consideration for in-game statistics (continued)

Variables in consideration	**Definitions**
Assists	An assist is awarded for a pass leading directly to a goal. [3]
Fouls	An illegal tackle by a player on an opponent resulting in a free kick, or in a penalty kick if the foul was adjudged to have been committed in the penalty area. [3]
Goalie Saves	Goalie Saves = Shouts on Goal -Scores [3]
Offside	To be in an offside position, a player must be on the opponent's half of the field & closer to the opponent's goal line than both the ball & the second-last defender. (However, the complete set of rules for offside is much more detailed.) The penalty for Offside is that an Indirect Free Kick is awarded to the opposing team to be taken from the place where the offside occurred. [5]

5.2. Model developed by using difference of seasonal averages

5.2.1. Develop models by using seasonal averages

All data was collected from NCAA.COM [6]. Seasonal averages were collected before the tournament started. For example, the first game of NCAA 2013 women's soccer tournament was held on November 15, 2013, the differences in seasonal averages were based on all games through November 10, 2013.

Data was collected for three years of the NCAA women's soccer tournament. This included the 2013, 2014 and 2015 tournaments. Seasonal averages for the variables listed in Table 5.1 (Set A) were collected for all the teams in the 2013 tournament. The variables included: Scoring Offense, Goals-Against Average, Shutout Percentage, Won-Lost-Tied Percentage, Save Percentage, Saves Per game, Assists Per Game and Points Per Game. Seasonal averages were also collected on the same variables for all teams playing each other in the 2014 and 2015 tournaments.

5.2.2. Develop models for the first round using seasonal averages

5.2.2.1. Develop ordinary least squares regression models

The response variable for the ordinary least squares regression model was point spread in the order of the team of interest minus the opposing team. Team of interest was the stronger team (higher seed numbers) in half of the games and the weaker team (lower seed numbers) in the other half of the games. A positive point spread indicates a win for the team of interest and a negative value indicates a loss for the team of interest. There were 192 teams playing 96 games in first rounds of the tournaments in 2013, 2014 and 2015. Tied games were excluded when developing the models. Differences of seasonal averages for 85 games were collected and considered in developing the models. For 42 games of the first round games in the three years, the point spread was obtained by using the stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the remainder of the games in the first rounds of the three years, the point spread was acquired by using the scores of weaker team (lower seed numbers) minus stronger team (higher seed numbers).

The intercept was excluded when developing the models because the models should give the same results regardless of the ordering of the teams in the model. Stepwise selection was used with an α value of 0.2 for both entry and exit to develop the models. The α value of 0.2 was used to allow more variables to enter the initial model since there are fewer variables in soccer to consider than in basketball and volleyball. The differences of the seasonal averages for all the variables previously given in Table 5.1 (Set A) between the two teams were considered for entry in the model.

5.2.2.1.1. Development of ordinary least squares regression model for the first round

The ordinary least squares regression model to help predict the winning team for each game in the first round based on using differences of seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-0.9076 * \text{Diff\_ Saves}) + (1.19331 * \text{Diff\_Assists})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 5.3. Table 5.4 gives the steps associated with the stepwise selection technique and Table 5.5 shows the associated R-square values as variables are added to the model. The model with the 2 significant variables explains an estimated 39% of the variation in point spread.

Table 5.3. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Saves	1	-0.90760	0.20255	-4.48	<.0001	1.28196
Assists	1	1.19331	0.39728	3.00	0.0035	1.28196

Table 5.4. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Saves		1	0.3258	0.3258	9.7235	40.59	<.0001
2	Assists		2	0.0661	0.3919	2.6326	9.02	0.0035

Table 5.5. Summary of R-squares value

Root MSE	2.57927	R-Square	0.3919
Dependent Mean	0.02353	Adj R-Sq	0.3772
Coeff Var	10962		

5.2.2.2. Develop logistic regression models

The logistic regression model was also fit to the data with the dependent variable recorded as ‘1’ for win and ‘0’ for loss for the team of interest. The model estimates the probability of a win for the team of interest. The team of interest was the stronger team (higher seed numbers) in half of the games and the weaker team (lower seed numbers) in the other half of the games.

No intercept was included during the development of the logistic regression model because the ordering of the teams in the model should not matter. Stepwise selection was used with an α value of 0.2 for both entry and exit when determining the significant variables in developing the logistic regression model. The differences of the seasonal averages for both teams for all previously mentioned variables in Table 5.1 (Set A) were considered for entry in the model.

5.2.2.2.1. Development of logistic regression model for the first round

A logistic regression model to help predict the winning team for each game in the first round was developed and found to be:

$$\pi(\text{Diff\_Saves, Diff\_Points})=\frac{e^{-0.7854*\text{Diff\_Saves}+0.3043*\text{Diff\_Points}}}{1+e^{-0.7854*\text{Diff\_Saves}+0.3043*\text{Diff\_Points}}}$$

Where π (Diff_Saves, Diff_Points) is the estimated probability that the team of interest will win the game with difference of seasonal averages in saves and difference of seasonal averages in points per game in model.

Table 5.6 shows the steps for the stepwise selection technique and Table 5.7 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 5.8 shows the Hosmer and Lemeshow Test [7] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.391 indicating that there was no evidence to reject using the logistic regression model.

Table 5.6. Summary of stepwise selection for logistic regression model

Step	**Effect**		**DF**	**Number**	**Score**	**Wald**	**Pr > ChiSq**
	Entered	**Removed**		**In**	**Chi-Square**	**Chi-Square**	
1	Saves		1	1	22.2778		<.0001
2	Points		1	2	4.7683		0.0290

Table 5.7. Logistic regression model parameter estimates

Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Saves	1	-0.7854	0.2397	10.7342	0.0011
Points	1	0.3043	0.1440	4.4670	0.0346

Table 5.8. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	**DF**	**Pr > ChiSq**
7.3736	7	0.3910

5.2.3. Develop models for the second round using seasonal averages

5.2.3.1. Develop ordinary least squares regression models

There were 96 teams playing 48 games in second rounds of the tournaments in 2013 to 2015. Tie games were excluded when developing the models. Differences of seasonal averages for 45 games were collected and considered for entry into the model. For 22 games of the second round, the point spread was obtained by using the scores of stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using stronger team (higher seed numbers) minus the weaker team (lower seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.2 for both entry and exit to develop the models. The differences of the seasonal averages of the previously mentioned variables listed in Table 5.1 (Set A) between the two teams were considered for entry in the model.

5.2.3.1.1. Development of ordinary least squares regression model for the second round

The ordinary least squares regression model to help predict the winning team for each game in the second round based on using differences of seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = 7.97536 * \text{Diff\_SavePct} + 1.53406 * \text{Diff\_Assists}$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 5.9. Table 5.10 gives the steps associated with the stepwise selection technique and Table 5.11 shows the associated R-square values as variables are added to the model. The model with the 2 significant variables explains an estimated 24% of the variation in point spread.

Table 5.9. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
SavePct	1	7.97536	4.40851	1.81	0.0774	1.09629
Assists	1	1.53406	0.42676	3.59	0.0008	1.09629

Table 5.10. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Assists		1	0.1814	0.1814	0.0530	9.75	0.0032
2	SavePct		2	0.0579	0.2393	-0.9921	3.27	0.0774

Table 5.11. Summary of R-squares value

Root MSE	2.29602	R-Square	0.2393
Dependent Mean	-0.26667	Adj R-Sq	0.2039
Coeff Var	-861.00722		

5.2.3.2. Develop logistic regression models

The logistic regression model was also fit for the data with responses recorded as‘1’ for win and ‘0’ for loss for the team of interest. No intercept was included during the development of the logistic regression model. Stepwise selection was used with an α value of 0.2 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences of the seasonal averages of all previously mentioned variables listed in Table 5.1 (Set A) between the two teams were considered for entry in the model.

5.2.3.2.1. Development of logistic regression model for the second round

A logistic regression model to help predict the winning team for each game in the second round was developed and found to be:

$$\pi(\text{Diff\_Assists})=\frac{e^{0.7953*\text{Diff\_Assists}}}{1+e^{0.7953*\text{Diff\_Assists}}}$$

Where π (Diff_Assists) is the estimated probability that the team of interest will win the game with difference of seasonal averages in assists in model.

Table 5.12 shows the steps for the stepwise selection technique and Table 5.13 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 5.14 shows the Hosmer and Lemeshow Test [7] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.2733 indicating that there was no evidence to reject using the logistic regression model.

Table 5.12. Summary of stepwise selection for logistic regression model

Step	**Effect**		**DF**	**Number**	**Score**	**Wald**	**Pr > ChiSq**
	Entered	**Removed**		**In**	**Chi-Square**	**Chi-Square**	
1	Assists		1	1	4.0025		0.0454

Table 5.13. Logistic regression model parameter estimates

Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Assists	1	0.7953	0.4190	3.6028	0.0577

Table 5.14. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	**DF**	**Pr > ChiSq**
8.7207	7	0.2733

5.2.4. Develop models for the third and higher rounds using seasonal averages

5.2.4.1. Develop ordinary least squares regression models

There were 90 teams playing 45 games in second rounds of the tournaments in 2011 to 2013. Tie games were not included when develop the models. Differences of seasonal averages for 38 games were collected and considered for entry into the model. For the first half games of the second round, the point spread was obtained by using the stronger team (higher seed numbers) minus the weaker team (lower seed numbers). For the remainder of games in the second round, the point spread was acquired by using the scores of weaker team (lower seed numbers) minus the stronger team (higher seed numbers). The intercept was excluded when developing the models. Stepwise selection was used with an α value of 0.2 for both entry and exit to develop the models. The differences of the seasonal averages of the previously mentioned variables listed in Table 5.1 (Set A) between the two teams were considered for entry in the model.

5.2.4.1.1. Development of ordinary least squares regression model for the third and higher rounds

The ordinary least squares regression model to help predict the winning team for each game in the third and higher rounds based on using differences between seasonal averages of the significant variables was developed and found to be:

$$\hat{Y} = (-7.2473*\text{Diff\_GoalsAgainst}) + (-6.50028*\text{ Diff\_Shutout\%}) + (-6.12538*\text{Diff\_WLT\%}) + (1.81157*\text{ Diff\_Assists})$$

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 5.15. Table 5.16 gives the steps associated with the stepwise selection technique and Table 5.17 shows the associated R-square values as variables are added to the model. The model with the 4 significant variables explains an estimated 68% of the variation in point spread.

Table 5.15. Point spread model parameter estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
GoalsAgainstAvg	1	-7.24730	1.47994	-4.90	<.0001	5.86470
ShutoutPct	1	-6.50028	2.41752	-2.69	0.0110	3.83753
WonLostTiedPct	1	-6.12538	2.76841	-2.21	0.0337	4.47427
AssistsPG	1	1.81157	0.42620	4.25	0.0002	2.64622

Table 5.16. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GoalsAgainstAvg		1	0.4497	0.4497	29.1979	30.23	<.0001
2	AssistsPG		2	0.1411	0.5908	14.4792	12.42	0.0012
3	ShutoutPct		3	0.0426	0.6334	11.4319	4.07	0.0514
4	WonLostTiedPct		4	0.0461	0.6795	7.9654	4.90	0.0337

Table 5.17. Summary of R-squares value

Root MSE	1.55333	R-Square	0.6795
Dependent Mean	-0.47368	Adj R-Sq	0.6418
Coeff Var	-327.92552		

5.2.4.2. Develop logistic regression models

The logistic regression model was also fit for the data with responses recorded as‘1’ for win and ‘0’ for loss for the team of interest. No intercept was used during the development of the logistic regression model. Stepwise selection was used with an α value of 0.2 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences of the seasonal averages of all previously mentioned variables listed in Table 5.1 (Set A) between the two teams were considered for entry in the model.

5.2.4.2.1. Development of logistic regression model for the third and higher rounds

A logistic regression model to help predict the winning team for each game in the third and higher rounds was developed and found to be:

$$\pi(\text{Diff\_GoalsAgainst, Diff\_Assists})=\frac{e^{-4.3317*\text{Diff\_GoalsAgainst}+1.8709*\text{Diff\_Assists}}}{1+e^{-4.3317*\text{Diff\_GoalsAgainst}+1.8709*\text{Diff\_Assists}}}$$

Where π (Diff_GoalsAgainst, Diff_Assists) is the estimated probability that the team of interest will win the game with difference of seasonal averages in goals against and difference of seasonal averages in assists in model.

Table 5.18 shows the steps for the stepwise selection technique and Table 5.19 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 5.20 shows the Hosmer and Lemeshow Test [7] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.2348 indicating that there was no evidence to reject using the logistic regression model.

Table 5.18. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	GoalsAgainstAvg		1	1	14.2971		0.0002
2	AssistsPG		1	2	7.7820		0.0053

Table 5.19. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
GoalsAgainstAvg	1	-4.3317	1.6588	6.8192	0.0090
AssistsPG	1	1.8709	0.7713	5.8842	0.0153

Table 5.20. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
10.4510	8	0.2348

5.2.5. Validating first round using models developed

5.2.5.1. Ordinary least squares regression model

The first ordinary least squares regression models developed by using seasonal averages was used to predict the first round of 2016 season to check the accuracy of the model. It is noted that the 2016 season was not used in the development of the models.

Table 5.21 gives the prediction results for the ordinary least squares regression model for first round of the NCAA 2016 women's soccer tournament.

Table 5.21. Accuracy of ordinary least squares regression model developed by seasonal averages when validating first round of 2016

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	7	7	14
	Loss	5	8	13
	Total	12	15	27
Overall Accuracy				55.56%

The first logistic regression model developed by using seasonal averages was used to predict the first round of 2016 season to check the accuracy of the model. It is noted that the 2016 season was not used in the development of the models.

Table 5.22 gives the prediction results for the logistic regression model for first round of the NCAA 2016 women's soccer tournament.

Table 5.22. Accuracy of logistic regression model developed by seasonal averages when validating first round of 2016

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	8	6	14
	Loss	4	9	13
	Total	12	15	32
Overall Accuracy				62.96%

It is noted that the percentage of accuracy for first rounds using the logistic regression models which developed by using seasonal averages is better than the ordinary least squares regression model.

5.2.6. Validating second round using models developed

The second round ordinary least squares regression model and logistic regression model that were developed by using seasonal averages were used to predict the second round of the 2016 soccer tournament to check the accuracy of the model. It is noted that the 2016 season was not used in the development of the models.

Table 5.23 gives the prediction results for the ordinary least squares regression model for second round of the NCAA 2016 women's soccer tournament. Table 5.24 gives equivalent results for the logistic regression model.

Table 5.23. Accuracy of ordinary least squares regression model developed by seasonal averages when validating second round of 2016

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	3	2	5
	Loss	1	8	9
	Total	4	10	14
Overall Accuracy				78.57%

Table 5.24. Accuracy of logistic regression model developed by seasonal averages when validating second round of 2016

Logistic		Predicted		
		Win	**Loss**	**Total**
Actual	**Win**	4	1	5
	Loss	2	7	9
	Total	6	8	14
Overall Accuracy				78.57%

5.2.7. Validating third and higher rounds using models developed

The third ordinary least squares regression model and logistic regression model developed by using seasonal averages were used to predict the third through final rounds of 2016 tournament to check the accuracy of the model.

Table 5.25 gives the validation results for the ordinary least squares regression model for third and higher rounds of the NCAA 2016 women's soccer tournament. Table 5.26 gives similar results for the logistic regression model.

Table 5.25. Accuracy of ordinary least squares regression model developed by seasonal averages when validating third and higher rounds of 2016

Point spread		Predicted		
		Win	**Loss**	**Total**
Actual	**Win**	2	4	6
	Loss	5	4	9
	Total	7	8	15
Overall Accuracy				40%

Table 5.26. Accuracy of logistic regression model developed by seasonal averages when validating third and higher rounds of 2016

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	1	5	6
	Loss	3	6	9
	Total	4	11	15
Overall Accuracy				46.67%

5.2.8. Bracketing the 2016 tournament before tournament begins – Prediction

Results were predicted for every round before the 2016 tournament began. Differences of significant seasonal averages of variables were found for all teams playing in the first round and put into first round model. Differences of seasonal averages for teams predicted to play each other in the second round were placed in second round model and winners of this round were predicted. Differences of seasonal averages of variables found to be significant of teams predicted to play each other in the third round were placed in the third round model and winning teams predicted for this round. This process continued until a tournament winner was selected.

The predicted results were then compared against the actual results for each round of games in the 2016 tournament.

5.2.8.1. Examples for each round in 2016 tournament

An example for the first round, second round and then third or higher round will be given as to how the ordinary least squares regression model for a particular round in 2016 tournament was used.

5.2.8.1.1. Ordinary least squares regression model developed by seasonal averages

5.2.8.1.1.1. Ordinary least squares regression model for first round

The ordinary least squares regression model for first round developed by using differences of seasonal averages is:

$$\hat{Y} = (-0.9076 * \text{Diff\_ Saves}) + (1.19331 * \text{Diff\_Assists})$$

Stanford played Houston Baptist in the first round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.27.

Table 5.27. Stanford and Houston Baptist Statistics

Team	Score	Saves*	Assists*
Stanford	4	2.42	1.84
Houston Baptist	0	5.1	1.33
Difference	4	-2.68	0.51

* Average per game for season

Using the model above, the game between Stanford and Houston Baptist had a predicted point spread of:

$\hat{y} = (-0.9076 * -2.68) + (1.19331 * 0.51) = 3.04$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Stanford, who won the game by a score of 4 to 0.

Rutgers played Harvard in the first round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.28.

Table 5.28. Rutgers and Harvard Statistics

Team	Score	Saves*	Assists*
Rutgers	3	3.33	1.38
Harvard	0	3.44	1.44
Difference	3	-0.11	-0.06

* Average per game for season

Using the model above, the game between Rutgers and Harvard had a predicted point spread of:

$\hat{y}$ = (-0.9076 * -0.11) + (1.19331* -0.06) = 0.03

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Rutgers, who won the game by a score of 3 to 0.

Utah played Texas Tech in the first round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.29.

Table 5.29. Utah and Texas Tech Statistics

Team	Score	Saves*	Assists*
Utah	1	3.95	1.47
Texas Tech	0	4.32	0.47
Difference	1	-0.37	1

* Average per game for season

Using the model above, the game between Utah and Texas Tech had a predicted point spread of:

$\hat{y}$ = (-0.9076 * -0.37) + (1.19331* 1) = 1.53

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Utah, who won the game by a score of 1 to 0.

Auburn played South Alabama in the first round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.30.

Table 5.30. Auburn and South Alabama Statistics

Team	Score	Saves*	Assists*
Auburn	4	3.15	2.2
South Alabama	0	4.1	1.48
Difference	4	-0.95	0.72

* Average per game for season

Using the model above, the game between Auburn and South Alabama had a predicted point spread of:

$\hat{y}$ = (-0.9076 *-0.95) + (1.19331*0.72) = 1.72

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Auburn, who won the game by a score of 4 to 0.

Round 1:

Number correct: 15

Number incorrect: 12

Total: 27

5.2.8.1.1.2. Ordinary least squares regression model for second round

The ordinary least squares regression model for second round developed by using differences of seasonal averages is:

$$\hat{Y} = (7.97536^* \text{ Diff\_SavePct}) + (1.53406^*\text{Diff\_Assists})$$

Rutgers played Georgetown in the second round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.31.

Table 5.31. Rutgers and Georgetown Statistics

Team	Score	Save_Pct*	Assists*
Rutgers	0	0.805	1.38
Georgetown	2	0.811	1.9
Difference	-2	-0.006	-0.52

* Average per game for season

Using the model above, the game between Rutgers and Georgetown had a predicted point spread of:

$\hat{y}$ = (7.97536* - 0.006) + (1.53406 * - 0.52) = -0.85

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Rutgers, who lost the game by a score of 0 to 2.

Wisconsin played Florida in the second round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.32.

Table 5.32. Wisconsin and Florida Statistics

Team	**Score**	**Save_Pct***	**Assists***
Wisconsin	2	0.795	1
Florida	3	0.753	2.4
Difference	-1	0.042	-1.4

* Average per game for season

Using the model above, the game between Wisconsin and Florida had a predicted point spread of:

$\hat{y}$ = (7.97536* 0.042) + (1.53406*(-1.4)) = -1.81

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Wisconsin, who lost the game by a score of 2 to 3.

Round 2:

Number correct: 7

Number incorrect: 7

Total: 14

5.2.8.1.1.3. Ordinary least squares regression model for third and higher rounds

The ordinary least squares regression model for third and higher rounds developed by using differences of seasonal averages is:

$$\hat{Y} = (-7.2473*\text{Diff\_GoalsAgainst}) + (-6.50028*\text{Diff\_Shutout\%}) + (-6.12538*\text{Diff\_WLT\%}) + (1.81157*\text{Diff\_Assists})$$

South Carolina played BYU in the third round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.33.

Table 5.33. South Carolina and BYU Statistics

Team	Score	Goals_Against*	Shutout%*	WLT%*	Assists *
South Carolina	1	0.441	0.55	0.925	1.95
BYU	0	0.464	0.632	0.868	2.89
Difference	1	-0.023	-0.082	0.057	-0.94

* Average per game for season

Using the model above, the game between South Carolina and BYU had a predicted point spread of:

$\hat{y} = (-7.2473*-0.023) + (-6.50028*-0.082) + (-6.12538*0.057) + (1.81157* -0.94) = -1.35$

Since $\hat{y} < 0$ this game was coded as an incorrectly predicted loss for South Carolina, who won the game by a score of 1 to 0.

Clemson played North Carolina in the third round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.34.

Table 5.34. Clemson and North Carolina Statistics

Team	Score	Goals_Against*	Shutout%*	WLT%*	Assists *
Clemson	0	0.774	0.35	0.725	2.65
North Carolina	1	0.666	0.45	0.75	1.55
Difference	-1	0.108	-0.1	-0.025	1.1

* Average per game for season

Using the model above, the game between Clemson and North Carolina had a predicted point spread of:

$\hat{y} = (-7.2473*0.108) + (-6.50028* -0.1) + (- 6.12538*-0.025) + (1.81157* 1.1) = 2.01$

Since $\hat{y} > 0$ this game was coded as an incorrectly predicted win for Clemson, who lost the game by a score of 0 to 1.

Round 3-6:

Number correct: 2

Number incorrect: 13

Total: 15

It should be noted that only the teams predicted to play each other in the second round were used in the model. The actual teams were not used all the time since predicting was done before the tournament started.

5.2.8.1.2. Logistic regression model developed by seasonal averages

An example for first round, second round and third or higher round will be given as to how the logistic regression model for a particular round was used for each round in 2016 tournament.

5.2.8.1.2.1. Logistic regression model for first round

The logistic regression model for first round developed by using differences of seasonal averages is:

$$\pi(\text{Diff\_Saves, Diff\_Points}) = \frac{e^{-0.7854*\text{Diff\_Saves}+0.3043*\text{Diff\_Points}}}{1+e^{-0.7854*\text{Diff\_Saves}+0.3043*\text{Diff\_Points}}}$$

Stanford played Houston Baptist in the first round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.35.

Table 5.35. Stanford and Houston Baptist Statistics

Team	Score	Saves*	Points*
Stanford	4	2.42	6.79
Houston Baptist	0	5.1	4.67
Difference	4	-2.68	2.12

* Average per game for season

Using the model above, the game between Stanford and Houston Baptist had an estimated probability of winning the game of:

$\pi(-2.68, 2.12) = \frac{e^{-0.7854*-2.68+0.3043*2.12}}{1+e^{-0.7854*-2.68+0.3043*2.12}} = 0.94$

Since π >0.5 this game was coded as a correctly predicted win for Stanford, who won the game by a score of 4 to 0.

Long Beach State played Santa Clara in the first round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.36.

Table 5.36. Long Beach State and Santa Clara Statistics

Team	Score	Saves*	Points*
Long Beach State	0	4	4.35
Santa Clara	3	3.47	3.26
Difference	-3	0.53	1.09

* Average per game for season

Using the model above, the game between Long Beach State and Santa Clara had an estimated probability of winning the game of:

$\pi(0.53, 1.09) = \frac{e^{-0.7854*0.53+0.3043*1.09}}{1+e^{-0.7854*0.53+0.3043*1.09}} = 0.48$

Since π <0.5 this game was coded as a correctly predicted loss for Long Beach State, who lost the game by a score of 0 to 3.

Virginia played Monmouth in the first round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.37.

Table 5.37. Virginia and Monmouth Statistics

Team	Score	Saves*	Points*
Virginia	4	2	6.53
Monmouth	1	2.85	8.45
Difference	3	-0.85	-1.92

* Average per game for season

Using the model above, the game between Virginia and Monmouth had an estimated probability of winning the game of:

π(-0.85, -1.92) = $\frac{e^{-0.7854*-0.85+0.3043*-1.92}}{1+e^{-0.7854*-0.85+0.3043*-1.92}}$ = 0.52

Since π >0.5 this game was coded as a correctly predicted win for Virginia, who won the game by a score of 4 to 1.

Albany played Connecticut in the first round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.38.

Table 5.38. Albany and Connecticut Statistics

Team	Score	Saves*	Points*
Albany	2	4.79	4.53
Connecticut	4	4.43	5.57
Difference	-2	0.36	-1.04

* Average per game for season

Using the model above, the game between Albany and Connecticut had an estimated probability of winning the game of:

π(0.36, -1.04) = $\frac{e^{-0.7854*0.36+0.3043*-1.04}}{1+e^{-0.7854*0.36+0.3043*-1.04}}$ = 0.35

Since π <0.5 this game was coded as a correctly predicted loss for Albany, who lost the game by a score of 2 to 4.

Round 1:

Number correct: 17

Number incorrect: 10

Total: 27

5.2.8.1.2.2. Logistic regression model for second round

The logistic regression model for second round developed by using differences of seasonal averages is:

π(Diff_Assists) = $\frac{e^{0.7953*\text{Diff\_Assists}}}{1+e^{0.7953*\text{Diff\_Assists}}}$

Stanford played Santa Clara in the second round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.39.

Table 5.39. Stanford and Santa Clara Statistics

Team	**Score**	**Assists***
Stanford	0	1.84
Santa Clara	1	0.84
Difference	-1	1

* Average per game for season

Using the model above, the game between Stanford and Santa Clara had an estimated probability of winning the game of:

$\pi(1)=\frac{e^{0.7953*1}}{1+e^{0.7953*1}} = 0.3$

Since π<0.5 this game was coded as a correctly predicted loss for Stanford, who lost the game by a score of 0 to 1.

Wisconsin played Florida in the second round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.40.

Table 5.40. Wisconsin and Florida Statistics

Team	**Score**	**Assists***
Wisconsin	2	1
Florida	3	2.4
Difference	-1	-1.4

* Average per game for season

Using the model above, the game between Wisconsin and Florida had an estimated probability of winning the game of:

$\pi(\text{-}1.4) = \frac{e^{0.7953*-1.4}}{1+e^{0.7953*-1.4}} = 0.45$

Since π<0.5 this game was coded as a correctly predicted loss for Wisconsin, who lost the game by a score of 2 to 3.

Round 2:

Number correct: 5

Number incorrect: 9

Total: 14

5.2.8.1.2.3. Logistic regression model for third and higher rounds

The logistic regression model for third and higher rounds developed by using differences of seasonal averages is:

$$\pi(\text{Diff\_GoalsAgainst, Diff\_Assists}) = \frac{e^{-4.3317*\text{Diff\_GoalsAgainst}+1.8709*\text{Diff\_Assists}}}{1+e^{-4.3317*\text{Diff\_GoalsAgainst}+1.8709*\text{Diff\_Assists}}}$$

Clemson played North Carolina in the third round of the 2016 tournament. Data on differences of significant seasonal averages was collected and displayed in Table 5.41.

Table 5.41. Clemson and North Carolina Statistics

Team	Score	Goals_Against*	Assists*
Clemson	0	0.774	2.65
North Carolina	1	0.666	1.55
Difference	-1	0.108	1.1

* Average per game for season

Using the model above, the game between Clemson and North Carolina had an estimated probability of winning the game of:

$$\pi(0.108, 1.1) = \frac{e^{-4.3317*0.108+1.8709*1.1}}{1+e^{-4.3317*0.108+1.8709*1.1}} = 0.03$$

Since $\pi < 0.5$ this game was coded as a correctly predicted loss for Clemson, who lost the game by a score of 0 to 1.

Round 3-6:

Number correct: 3

Number incorrect: 12

Total: 15

It should be noted that only the teams predicted to play each other in the second round were used in the model. The actual teams were not used all the time since predicting was done before the tournament started.

5.2.9. Results for Prediction by using models developed by difference of seasonal averages

In 2016, a continuous process was used in verifying the models instead of doing round by round predictions as in previous chapter. In other words, a complete bracket was filled out in 2016 before any game was played.

The ordinary least squares regression model for the first round developed by using difference of seasonal averages were used to predict the teams who go to next round. Once the teams in the second round were predicted, the second-round models were used to predict the winners of the second round. This process was continued for the third and higher rounds until the predicted final winner of the game was determined.

Accuracy of ordinary least squares regression model results are given in Table 5.42 and results of logistic regression models are given in Table 5.43.

Table 5.42. Prediction results of each round for 2016: (Ordinary least squares regression model developed by seasonal averages)

	Correct	**Incorrect**	**Total games**
First round	15	12	27
Second round	7	7	14
Third round	2	6	8
Fourth round	0	4	4
Fifth round	0	2	2
Final round	0	1	1
	Overall Accuracy		42.86%

Table 5.43. Prediction results of each round for 2016: (Logistic regression model developed by seasonal averages)

	Correct	Incorrect	Total games
First round	17	10	27
Second round	5	9	14
Third round	2	6	8
Fourth round	1	3	4
Fifth round	0	2	2
Final round	0	1	1
	Overall Accuracy		44.64%

It is noted logistic regression model works slightly better than ordinary least squares regression model when using seasonal averages to develop models.

5.3. Model developed by using difference of in-game statistics

5.3.1. Develop models by using in-game statistics

Data was collected for NCAA women's soccer tournament of 2016. Tie games were excluded when develop the models. In-game statistics were collected for 55 games of 63 games of the 2015 tournament on the variables listed in Table 5.2 (Set B). The variables included: Shutout, SOG, Assists, Fouls, Goalie Saves and Offside.

5.3.1.1. Develop ordinary least squares regression model using in-game statistics

The response variable for the ordinary least squares regression model was point spread in the order of the team of interest minus the opposing team. Team of interest was the stronger team (higher seed numbers) in half of the games and the weaker team (lower seed numbers) in the other half of the games. A positive point spread indicates a win for the team of interest and a negative value indicates a loss for the team of interest. There were 64 teams playing 63 games in the tournaments of 2015. However, only 55 games were left after eliminating the tie games.

The intercept was excluded when developing the models because the models should give the same results regardless of the ordering of the teams in the model. Stepwise selection was used with an α value of 0.2 for both entry and exit to develop the models. The differences of the seasonal averages for all the variables previously given in Table 5.2 (Set B) between the two teams were considered for entry in the model.

The ordinary least squares regression model to help predict the winning team for each round based on using differences between in-game statistics of significant variables was developed and found to be:

$$\hat{Y} = (0.99212 * \text{Diff\_SOG}) + (0.99038 * \text{Diff\_GoaliesSaves})$$

The following statistics have positive coefficients associated with them which is to be expected: Difference in SOG and Difference in Goalies Saves. It is noted that each additional Goalies Saves over the opposing team is estimated to be worth on average approximately 1 point. Each additional SOG over the other team is worth approximately 1 point.

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 5.44. Table 5.45 gives the steps associated with the stepwise selection technique and Table 5.46 shows the associated R-square values as variables are added to the model. The model with the 2 significant variables explains an estimated 99% of the variation in point spread.

Table 5.44. Point spread model parameter estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
SOG	1	0.99212	0.01721	57.63	<.0001	8.40089
GoaliesS	1	0.99038	0.02239	44.24	<.0001	8.40089

Table 5.45. Summary of stepwise selection for point spread model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	A		1	0.7056	0.7056	1111.59	129.42	<.0001
2	SOG		2	0.0824	0.7880	787.547	20.61	<.0001
3	GoaliesS		3	0.1993	0.9874	1.0312	819.54	<.0001
4		A	2	0.0000	0.9874	-0.9678	0.00	0.9751

Table 5.46. Summary of R-squares value

Root MSE	0.27418	R-Square	0.9874
Dependent Mean	0.23636	Adj R-Sq	0.9869
Coeff Var	115.99884		

5.3.1.2. Develop logistic regression model using in-game statistics

The logistic regression model was also fit for the data with responses recorded as'1' for win and '0' for loss for the team of interest. No intercept was used during the development of the logistic regression model. Stepwise selection was used with an α value of 0.2 for both entry and exit when determine the significant variables in developing the logistic regression model. The differences of the seasonal averages of all previously mentioned variables listed in Table 5.2 (Set B) between the two teams were considered for entry in the model.

A logistic regression model to predict the winning team for each round was developed and found to be:

$$\pi(\text{Diff\_SOG, Diff\_GoaliesS})=\frac{e^{3.254*\text{Diff\_SOG}+3.2308*\text{Diff\_GoaliesS}}}{1+e^{3.254*\text{Diff\_SOG}+3.2308*\text{Diff\_GoaliesS}}}$$

Where π (Diff_SOG, Diff_GoaliesS) is the estimated probability that the team of interest will win the game with difference of in-game statistics in SOG and difference of in-game statistics in Goalies saves in model.

Table 5.47 shows the steps for the stepwise selection technique and Table 5.48 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 5.49 shows the Hosmer and Lemeshow Test [7] was done to test whether there was evidence the logistic regression model was not appropriate. The p-value was 0.7568 indicating that there was no evidence to reject using the logistic regression model.

Table 5.47. Summary of stepwise selection for logistic regression model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	**Removed**		In	Chi-Square	Chi-Square	
1	A		1	1	22.7008		<.0001
2	SOG		1	2	3.9977		0.0456
3	GoaliesS		1	3	20.7524		<.0001
4		A	1	2		0.0743	0.7852

Table 5.48. Logistic regression model parameter estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
SOG	1	3.2540	0.9991	10.6071	0.0011
GoaliesS	1	3.2308	1.0381	9.6858	0.0019

Table 5.49. Hosmer and Lemeshow Goodness-of-Fit test

Chi-Square	DF	Pr > ChiSq
4.1972	7	0.7568

5.3.2. Validating 2015 first round using models developed

5.3.2.1. Verification of the models developed by using in-game statistics

Using the ordinary least squares regression model developed for the whole tournament, the point spread of each of 63 games in the 2015 tournament was estimated.

To verify the accuracy of the results for the ordinary least squares regression model, significant differences of in-game statistics were placed in the model developed.

The estimated response $\hat{y}$ then observed. If $\hat{y}$ is great than 0, a predicted win for the team of interest was coded. If $\hat{y}$ is less than 0, a predicted loss for the team of interest was coded.

To verify the accuracy of the results for the logistic regression model for the first round, a similar process was conducted. For each round of the game, statistics for the significant factors were collected and the difference was taken and placed into the logistic regression model to find a predicted probability, $\pi\ x_i$. If π xi was greater than 0.5, a predicted win was coded. If π xi was less than 0.5, a predicted loss was coded.

The second round and higher round models were verified in a similar way. Once the teams in the second round were determined, the same model was used to predict the winners of the second round. This process continued for the third and higher rounds.

The validation results for first round using ordinary least squares regression model and logistic regression model developed by using in-game statistics are given in Table 5.50 and Table 5.51, respectively.

Table 5.50. Accuracy of ordinary least squares regression model developed by in-game statistics when validating first round of 2015

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	13	0	13
	Loss	1	15	16
	Total	14	15	29
Overall Accuracy				96.55%

Table 5.51. Accuracy of logistic regression model developed by in-game statistics when validating first round of 2015

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	13	0	13
	Loss	1	15	16
	Total	14	15	29
Overall Accuracy				96.55%

It is noted that the percentage of accuracy for first round using the ordinary least squares model and using the logistic regression model are the same.

5.3.3. Validating second round using models developed

The validation results for second round using ordinary least squares regression model and logistic regression model developed by using in-game statistics are given in Table 5.52 and Table 5.53, respectively.

Table 5.52. Accuracy of ordinary least squares regression model developed by in-game statistics when validating second round of 2015

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	8	0	8
	Loss	0	7	7
	Total	8	7	15
Overall Accuracy				100%

Table 5.53. Accuracy of logistic regression model developed by in-game statistics when validating second round of 2015

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	8	0	8
	Loss	0	7	7
	Total	8	7	15
Overall Accuracy				100%

It is noted that the percentage of accuracy for second round using the ordinary least squares model and using the logistic regression model are equivalent.

5.3.4. Validating third and higher rounds using models developed

The validation results for third and higher rounds using ordinary least squares regression model and logistic regression model developed by using in-game statistics are given in Table 5.54 and Table 5.55, respectively.

Table 5.54. Accuracy of ordinary least squares regression model developed by in-game statistics when validating third and higher rounds of 2015

Point spread		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	6	0	6
	Loss	0	6	6
	Total	6	6	12
Overall Accuracy				100%

Table 5.55. Accuracy of logistic regression model developed by in-game statistics when validating third and higher rounds of 2015

Logistic		**Predicted**		
		Win	**Loss**	**Total**
Actual	**Win**	6	0	6
	Loss	0	6	6
	Total	6	6	12
Overall Accuracy				100%

It is noted that the percentage of accuracy for third and higher rounds using the logistic regression model is the same as the ordinary least squares regression model.

5.3.5. Bracketing the 2016 tournament before tournament begins – Predicting

For predicting, since the in-game statistics will not be available before the tournament, seasonal averages of the current year were collected and put into the in-game model to predict the winners of the soccer games for 2016 tournament.

Results were predicted for each round by using the ordinary least squares regression model and the logistic regression model developed based on in-game statistics before the 2016 tournament began.

Differences of seasonal averages of in-game statistics found to be significant were put into first round model based on the team playing. Differences of seasonal averages for each team predicted to play each other in the second round were placed in second round model and winners of this round were predicted. Differences of seasonal averages of variables of teams predicted to play each other in the third round were placed in the third round model and winning teams predicted for this round. This process continued until a winner is selected.

The predicted results were then compared against the actual results for each round of the tournament for 2016.

5.3.5.1. Examples for first round in 2016 tournament

An example for the first round, second round and then third or higher round will be given as to how the ordinary least squares regression model for a particular round in 2016 tournament was used.

5.3.5.1.1. Using ordinary least squares regression model developed by in-game statistics

The ordinary least squares regression model for first through final rounds developed by using differences of in-game statistics is:

$$\hat{Y} = (0.99212 * \text{Diff\_SOG}) + (0.99038 * \text{Diff\_GoaliesSaves})$$

USC played Eastern Washington in the first round of the 2015 tournament. Data on significant differences of in-game statistics were collected and displayed in Table 5.56.

Table 5.56. USC and Eastern Washington Statistics

Team	Score	SOG*	Goalies_Saves *
USC	3	4.12	6.94
Eastern Washington	1	6.29	3.81
Difference	2	-2.17	3.13

* Average per game for season

Using the model above, the game between USC and Eastern Washington had a predicted point spread of:

$\hat{y}$ = (0.99212* -2.17) + (0.99038 *3.13) = 0.95

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for USC, who won the game by a score of 3 to 1.

Texas A&M played TCU in the first round of the 2015 tournament. Data on significant differences of in-game statistics were collected and displayed in Table 5.57.

Table 5.57. Texas A&M and TCU Statistics

Team	Score	SOG*	Goalies_Saves *
Texas A&M	1	6.9	3.8
TCU	0	6.15	3.85
Difference	1	0.75	-0.05

* Average per game for season

Using the model above, the game between Texas A&M and TCU had a predicted point spread of:

$\hat{y}$ = (0.99212* 0.75) + (0.99038 * -0.05) = 0.69

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Texas A&M, who won the game by a score of 1 to 0.

Round 1:

Number correct: 11

Number incorrect: 16

Total: 27

5.3.5.1.2. Using logistic regression model developed by in-game statistics

An example for first round, second round, third or high round will be given as to how the logistic regression model for a particular round was used for each round in 2016 tournament.

The logistic regression model for first through final rounds developed by using differences of in-game statistics is:

$$\pi(\text{Diff\_SOG}, \text{Diff\_GoaliesS}) = \frac{e^{3.254*\text{Diff\_SOG}+3.2308*\text{Diff\_GoaliesS}}}{1+e^{3.254*\text{Diff\_SOG}+3.2308*\text{Diff\_GoaliesS}}}$$

USC played Eastern Washington in the first round of the 2015 tournament. Data on significant differences of in-game statistics were collected and displayed in Table 5.58.

Table 5.58. USC and Eastern Washington Statistics

Team	**Score**	**SOG***	**Goalies_Saves ***
USC	3	4.12	6.94
Eastern Washington	1	6.29	3.81
Difference	2	-2.17	3.13

* Average per game for season

Using the model above, the game between USC and Eastern Washington had an estimated probability of winning the game of:

$$\pi(-2.17, 3.13) = \frac{e^{3.254*-2.17+3.2308*3.13}}{1+e^{3.254*-2.17+3.2308*3.13}} = 0.09$$

Since $\pi < 0.5$this game was coded as an incorrectly predicted loss for USC, who won the game by a score of 3 to 1.

Texas A&M played TCU in the first round of the 2015 tournament. Data on significant differences of in-game statistics were collected and displayed in Table 5.59.

Table 5.59. Texas A&M and TCU Statistics

Team	Score	SOG*	Goalies_Saves *
Texas A&M	1	6.9	3.8
TCU	0	6.15	3.85
Difference	1	0.75	-0.05

* Average per game for season

Using the model above, the game between Texas A&M and TCU had an estimated probability of winning the game of:

$$\pi(0.75, -0.05) = \frac{e^{3.254*0.75+3.2308*-0.05}}{1+e^{3.254*0.75+3.2308*-0.05}} = 0.21$$

Since π <0.5 this game was coded as an incorrectly predicted loss for Texas A&M, who won the game by a score of 1 to 0.

Round 1:

Number correct: 11

Number incorrect: 16

Total: 27

5.3.6. Results for prediction by using models developed by in-game statistics

Ideally, a complete bracket was filled out in 2016 by using the ordinary least squares regression model and logistic regression model developed by using in-game statistics before any game was played. However, after predicting the first round of 2016, the accuracy of predicting was low. In other word, put seasonal averages into in-game model did not work well.

The Ordinary least squares regression model that developed by using in-game statistics was used to predict the team who go to next round, the results are given in Table 5.60.

A similar process was conducted for logistic regression model. However, the results were similar with ordinary least squares regression model, the results are given in Table 5.61.

Table 5.60. Prediction results of first round for 2016: (Ordinary least squares regression model developed by in-game statistics)

	Correct	Incorrect	Total games
First round	11	16	27
	Accuracy		40.74%

Table 5.61. Prediction results of first round for 2016: (Logistic regression model developed by in-game statistics)

	Correct	Incorrect	Total games
First round	11	16	32
	Accuracy		40.74%

Accuracy of both ordinary least squares regression model and logistic regression model has an accuracy of 41% since seasonal averages data not fit the in-game model well.

5.4. Conclusion

5.4.1. Validation - Models developed by using seasonal averages

To verify the accuracy of prediction results for the ordinary least squares regression model, differences of the seasonal averages for both teams for all significant variables were placed in the model developed for the whole tournament. The ordinary least squares regression model and the logistic regression model developed by using difference in seasonal averages for the first round had approximately a 55.56% and a 62.96% chance of correctly predicting the results, respectively. The ordinary least squares regression model and the logistic regression model developed by using difference in seasonal averages for the second round both had approximately a 78.57% chance of correctly predicting the results. The ordinary least squares regression model and the logistic regression model developed by using difference of seasonal averages for the third and higher rounds had approximately a 40% and 46.67% chance of correctly predicting the results, respectively.

5.4.2. Prediction - Models developed by using seasonal averages

In 2016, a continuous process was used in verifying the models instead of doing round by round predictions as in 2015. Namely, a complete bracket was filled out in 2016 before any game was played. When the differences of the seasonal averages for both teams for all significant variables were considered for entry in the ordinary least squares models and the logistic regression models which developed by using seasonal average, the models had approximately a 47.6% chance of correctly predicting the winner of a soccer game, respectively.

5.4.3. Validation - Models developed by using in-game statistics

To verify the accuracy of prediction results for the ordinary least squares regression model developed by using in-game statistics, differences of the in-game statistics for both teams for significant variables were placed in the model developed for the whole tournament. The ordinary least squares regression model and the logistic regression model for the first round both had a 96.55% chance of correctly predicting the results when the tie games were excluded. The ordinary least squares regression model and the logistic regression model for the second through final rounds both had a 100% chance of correctly predicting the results when the tie games were excluded.

5.4.4. Prediction - Models developed by using in-game statistics

When the differences of the seasonal averages were placed into the model developed by using differences of in-game statistics, the ordinary least squares regression model and logistic regression model both had approximately a 41% chance of correctly predicting for the winner of a soccer game for the first round.

It is noted that the predictions were done and brackets filled out before the tournament began. The accuracy is lower because tams predicted to play in the second round or higher rounds might not have actually made it to those rounds.

5.4.5. Overall comparisons

Both the ordinary least squares regression model and the logistic regression model developed by using in-game statistics work well when the in-game statistics are known.

However, when predicting results for future tournaments without in-game statistics given, the results are not good due the limited access of data readily available. Logistic model developed by using seasonal average with an overall 45% accuracy works better than the models developed by using in-game statistics. This is not surprising since the model developed using seasonal averages is developed with seasonal averages in mind. The model developed using in-game statistics is not, and then replacing in-game statistics with seasonal averages.

In order to improve this accuracy in the future, perhaps additional seasonal average variables and in-game statistics could be found which help to further explain the point margin in a women's soccer game.

5.5. References

[1] NCAA Division I Women's Soccer Championship. Retrieved October 01, 2017, from https://en.wikipedia.org/wiki/NCAA_Division_I_Women's_Soccer_Championship

[2] Road to the Championship. Retrieved October 01, 2017, from http://www.ncaa.com/championships/soccer-women/d1/road-to-the-championship

[3] Official Soccer Statistics Rules. Retrieved October 10, 2017, from http://fs.ncaa.org/Docs/stats/Stats_Manuals/Soccer/2009ez.pdf

[4] Soccer Glossary of Statistics. Retrieved October 10, 2017, from http://www.hometeamsonline.com/teams/popups/Glossary.asp?s=soccer

[5] SOCCER STATISTICAL DEFINITIONS. Retrieved October 10, 2017, from http://nmsoccer.com/users/nms/Data/soccerstatisticaldefinitions.pdf

[6] Ranking Summary. Retrieved June 10, 2017, from http://web1.ncaa.org/stats/StatsSrv/ranksummary

[7] Hosmer, David W.; Lemeshow, Stanley (2013). *Applied Logistic Regression*. New York: Wiley. ISBN 978-0-470-58247-3