BRACKETING THE NCAA WOMEN'S BASKETBALL TOURNAMENT

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

By

Wenting Wang

In Partial Fulfillment
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

May 2014

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Bracketing the NCAA Women's Basketball Tournament

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

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Rhonda Magel
Chair

Dr. Gang Shen

Dr. Frank Manthey

Approved:

5/9/2014
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

This paper presents a bracketing method for all the 63 games in NCAA Division I Women's basketball tournament. Least squares models and logistic regression models for Round 1, Round 2 and Rounds 3-6 were developed, to predict winners of basketball games in each of those rounds for the NCAA Women's Basketball tournament. For the first round, three-point goals, free throws, blocks and seed were found to be significant; For the second round, field goals and average points were found to be significant; For the third and higher rounds, assists, steals and seed were found to be significant. A complete bracket was filled out in 2014 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 least squares models, the models had approximately a 76% chance of correctly predicting the winner of a basketball game.

ACKNOWLEDGMENTS

I express my deepest appreciation and thanks to my major adviser, Dr. Rhonda Magel, for her guidance, encouragement, support, and assistance in this study and the suggestions she made in reviewing the original manuscript.

My appreciation and thanks are also expressed to my graduate committee, Dr. Gang Shen and Dr. Frank Manthey, for their assistance and review of this manuscript.

I express my thanks and appreciation to all the committee members in Department of Statistics for providing the opportunity to study in the Department of Statistics at North Dakota State University.

I am grateful to my father and mother and my father-in-law and mother-in-law and my husband for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1. INTRODUCTION

1.1. The History of NCAA Women's Division I Basketball Tournament

Women's basketball is becoming more and more popular, spreading from the east coast of the United States to the west coast, in large part among women's colleges. The National Collegiate Athletic Association (NCAA) Women's Division I basketball Tournament is an annual college basketball tournament for women. The Tournament is held each spring from March to April in all neutral venues. The Women's Championship was inaugurated in the 1981-1982 season. The NCAA tournament was preceded by the Association for Intercollegiate Athletics for Women's basketball (AIAW), which was held every year from 1972 to 1982. In 1982, both tournaments co-existed in a competitive way, rather than in parallel way. One year later, NCAA won the battle and AIAW disbanded.

College basketball has such national popularity and interest in the Women's Division I Championship have grown over these years. In 2003, the final championship game was moved to the Tuesday following the Monday men's championship game. Before 2003, the Women's Final Four was usually played before the men's Final Four. This means the women's championship game is now the final overall game of the college basketball season.

Unlike the men's tournament, there is no play-in game for women's tournament. There are a total of 64 qualified teams to play in March and April, 31 of which can earn automatic bids by winning their respective conference tournaments. The remaining teams are granted "at-large" bids, which are extended by the NCAA Selection Committee. The tournament is split into four regional tournaments- Midwest, West, East and South Regional, and each Regional has teams seeded from 1 to 16. The top-seeded team in each Regional plays with the 16th team, the second-

ranked team plays with the 15th, *etc*. Figure 1 shows the 2014 NCAA Women's basketball tournament bracket.

1.2. The Playing Rule and Structure

The women's tournament, like the men's tournament, is staged in a single elimination format which is also called an Olympic system. In other words, the loser of each game or bracket is immediately eliminated from winning the championship in the event. This format is part of the media and public frenzy known colloquially as *March Madness* or *The Big Dance*.

There are six rounds of the tournament in each season so there will be 63 games in total. The six rounds are Round64, Round32, Sweet16, Elite8, Final4 and Championship, respectively. There are 64 teams to play 32 games in Round64; 32 teams to play 16 games in Round32; 16 teams to play 8 games in Sweet16; 8 teams to play 4 games in Elite8; 4 teams play 2 games in Final4 and 2 teams battle the Championship. The current NCAA Women's Division I Basketball tournament structure in recent seasons is illustrated in Figure 2.

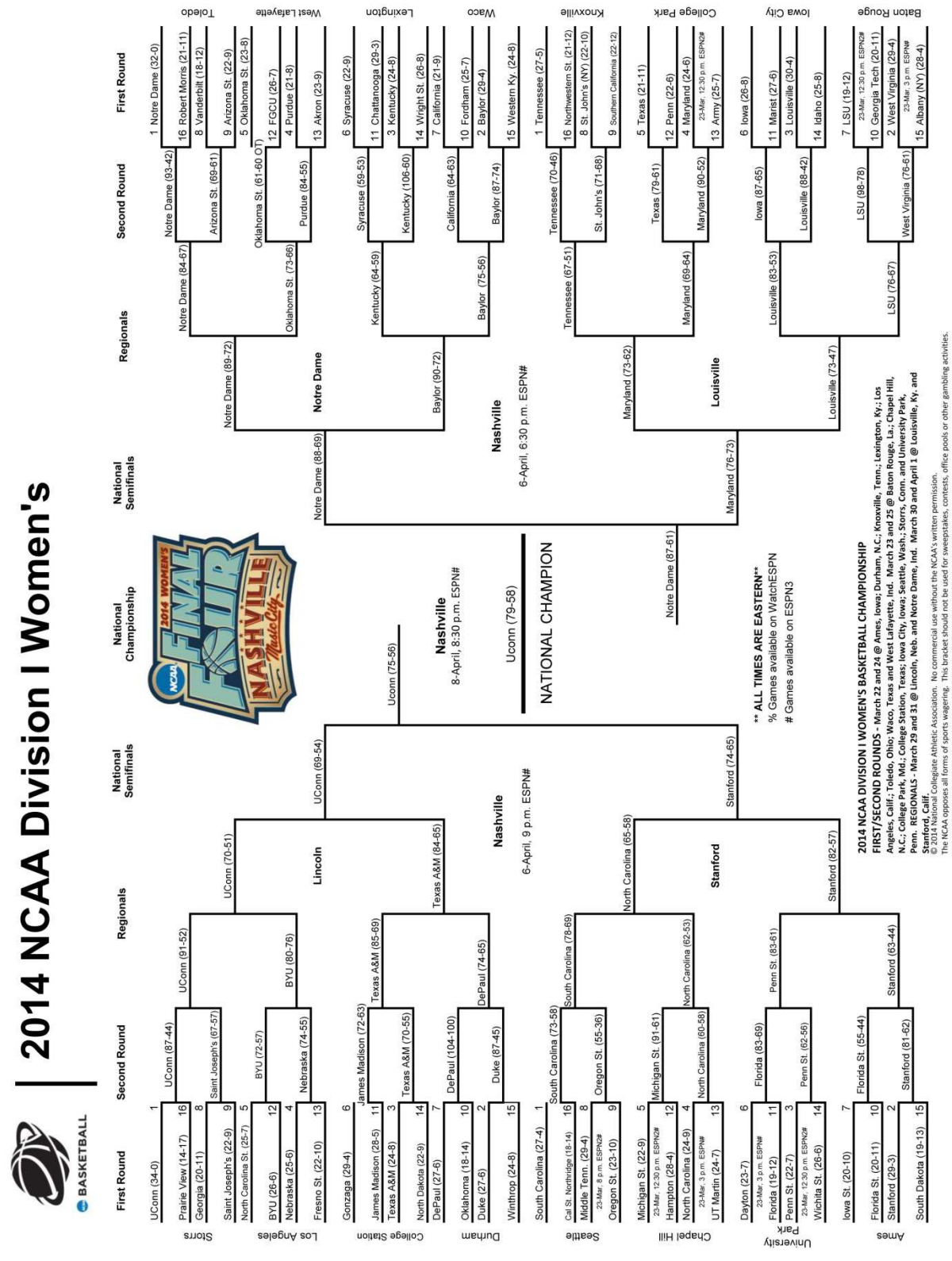

Figure 1: The NCAA women's basketball tournament bracket for the 2013 – 2014 season (This bracket is downloaded from: http://www.ncaa.com/interactive-bracket/basketball-women/d1)

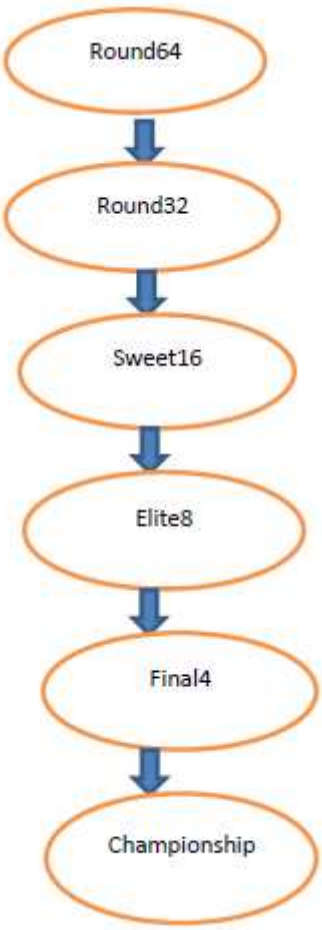

Figure 2: Structure of NCAA women's division I basketball tournament

CHAPTER 2. REVIEW OF PAST STUDIES

It is hard to find the articles related to predicting NCAA women's basketball game. Research has consistently shown that men's sports draw dominantly more attention than women's sports games, even though it is reportedly shown that there was a significant increase in the number of women or girls who actively participate or regularly play organized sports games by certain associations. (Kane, 1996; Duncan, 2006).

Previous works regarding the topic of factors affecting men's basketball games were reviewed, and some are mentioned here.

Carlin (1996) used very basic regression models to predict probability of winning using seed positions and computer ranking.

Schwertman, Schenk and Holbrook (1996) modified the approach to fit linear and logistic regression models for P(i,j) as a function of the difference in either team seeds or normal scores of the seeds on the basis of the data from 600 games (1985-1994).

Smith and Schwertman (1999) conducted an interesting research from a different angle to accurately predict the actual margin of victory upon the development of more complex regression models with the use of the seed position information. They used PRESS, which directly measures the predictive quality of a model, to determine the subset of independent variables that comprise the best prediction model. They proposed the following model for predicting the point spread of a men's basketball game in the NCAA tournament:

$$\hat{y} = -2.148X_1 + 1.68X_2 - .179X_1 X_2 + .260X_1^2 + .197X_{16}$$

Where $\hat{y}$ is the predicted margin of victory, X_1 is the lower seed numbers, X_2 is the higher seed numbers, and $X_{16}$is the linear yearly trend.

Caudill (2003) also used seed values and developed the maximum score estimator in the case of the NCAA men's basketball tournament to predict winnings. It was found that use of the maximum score estimator yielded slightly better results than results obtained through use of the probit/maximum likelihood models.

Kubatko, Oliver, Pelton and Rosenbaum (2007) proposed a good starting point for future basketball research. They analyzed the possession concept and found it to be connected with various statistics. Other important concepts have been included in their study, such as offensive and defensive ratings, plays, per-minute statistics, pace adjustments, true shooting percentage, effective field goal percentage and rebound rates.

West (2006) used a rating method based on ordinal logistic regression and expectation (the OLRE method) to predict the probability of winning for a basketball team playing in the NCAA Men's basketball tournament. West (2006) estimated the probabilities that a given team, i, would win 0 games, 1 games, 2 games, through 6 games.

Zhang (2013) used data form the 2002 -2012 season of NCAA Men's basketball tournament as the training data and then tested the accuracy for bracketing all 63 games in the 2012-2013 season. This study focused on bracketing the NCAA Men's basketball tournament by use of a conditional logistic probability model. This work is a modification of the work of West (2006). It was found that the conditional logistic probability model outperformed the restricted OLRE model proposed by West (2006) for 2013 March Madness.

Magel and Unruh (2013) analyzed NCAA Men's basketball games and found four common statistics were significant to determine winning, i.e. assists, free throw attempts, defensive rebounds and turnovers. Two models were developed by the use of a random sample

of 150 games chosen from 2009-2010 season and the 2010-2011 season. The models were used to bracket 2013 March Madness and correctly predicted 62% and 68% of the game results.

CHAPTER 3. DESCRIPTION OF STUDY

3.1. Research Objectives

The research objectives for this study include the following:

1) Develop least squares regression models for Round 1, Round 2 and Rounds 3-6, to predict winners of basketball games in each of those rounds for the NCAA Women's Basketball tournament; and
2) Develop Logistic regression models for Round 1, Round 2 and Rounds 3-6, to predict winners of basketball games in each of those rounds for the NCAA Women's Basketball tournament.

Data was collected for two seasons of the NCAA Women's Basketball tournament. This included the 2011 and 2012 tournaments. Seasonal averages were collected for all the teams in the 2011 tournament on the following variables: Field Goal Percentage; 3-pt Goal Percentage; Free Throws Percentage; Number of Rebounds; Number of Assists; Number of Blocks; Number of Steals and Average number of points. Seasonal averages were also collected on the same variables for all teams playing in the 2012 tournament. The seed number that each team was given in either the 2011 or 2012 tournament was also noted.

Two groups of models were developed by using the data collected from the two seasons. The first group of models used least squares regression with point spread as a response, and the second groups of models used a logistic regression approach with responses recorded as '1' for win and '0' for loss.

3.2. Develop Models for the First Round Using 2011 and 2012 Data

3.2.1. Develop Least Squares Regression Models

The response variable for the 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 128 teams playing 64 games in first rounds of the tournaments in 2011 and 2012. For the 32 games of the first round in 2011, 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 32 games of the first round of the tournaments in 2012, 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 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 0.15 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 were considered for entry in the model. The differences between seeds were also considered.

The generalized least squares model will be $y = x\beta + \varepsilon$, where y is the point spread, x is the matrix consisting of independent significant factors, β is the vector of coefficients corresponding to the independent factors, and ε is the random error.

3.2.2. Development of 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 logistic regression model

will be $\pi x_i = \frac{e^{x_i'\beta}}{1+ e^{x_i'\beta}}$ where $x_i'\beta = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ (Abraham & Ledolter, 2006) with πx_i estimating the probability of a win for the team of interest.

No intercept will be used during the development of the logistic model because the ordering of the teams in the model should not matter. Stepwise selection was used with an α value of 0.15 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 were considered for entry in the model. The differences between seeds were also considered for entry into the model.

3.3. Develop Models for the Second Round Using 2011 and 2012 Data

3.3.1. Develop Least Squares Regression Models

There were 64 teams playing 32 games in second rounds of the tournaments in 2011 and 2012. For the 16 games of the second round in 2011, 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 16 games of the second round in 2012, 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 0.15 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. The differences between seeds were also considered.

The generalized least squares model will be $y = x\beta + \varepsilon$, where y is the point spread, x is the matrix consisting of independent significant factors, β is the vector of coefficients corresponding to the independent factors, and ε is the random error.

3.3.2. Development of 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. The logistic regression model will be $\pi x_i = \frac{e^{x_i'\beta}}{1+ e^{x_i'\beta}}$ where $x_i'\beta = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ (Abraham & Ledolter, 2006) with πx_i estimating the probability of a win for the team of interest. No intercept will be used during the development of the logistic model. Stepwise selection was used with an α value of 0.15 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 were considered for entry in the model. The differences between seeds were also considered.

3.4. Develop Models for the Third and Higher Rounds Using 2011 and 2012 Data

3.4.1. Develop Least Squares Regression Models

There were 60 teams playing 30 games in third and higher rounds of the tournaments in 2011 and 2012. For the 15 games of the third and higher rounds in 2011, 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 15 games of the third and higher rounds in 2012, the point spread was got 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 0.15 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. The differences between seed values were also considered.

The generalized least squares model will be $y = x\beta + \varepsilon$, where y is the point spread, x is the matrix consisting of independent significant factors, β is the vector of coefficients corresponding to the independent factors, and ε is the random error.

3.4.2. Development of Logistic Regression Model

The logistic regression model was also fit for the data with responses recorded as a '1' for win and '0' for a loss for the team of interest. The logistic regression model will be $\pi x_i = \frac{e^{x_i'\beta}}{1+ e^{x_i'\beta}}$ where $x_i'\beta = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ (Abraham & Ledolter, 2006) with πx_i estimating the probability of a win for the team of interest. No intercept will be used when developing the logistic model. Stepwise selection was used with an α value of 0.15 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 the previously mentioned variables were considered for entry in the model. Differences between seeds was also considered.

3.5. Verification of the Models

Using the least squares regression model developed for the first round, the point spread of 16 games in the first round of the 2013 tournament was estimated based of the stronger perceived team (higher seed number). The point spread for the remaining 16 games of round 1 was estimated based on the team with higher seed number minus team with lower seed value.
To verify the accuracy of prediction results for the least squares regression model, values of variables were placed in the model developed for the first round. The estimated response $\hat{y}$ then observed.

If $\hat{y} > 0$, a predicted win for the point spread model was coded.

If $\hat{y} < 0$, a predicted loss for the point spread model 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, statistics for the significant factors were collected and the difference was taken and placed into the logistic models to find a predicted probability, πx_i.

If $\pi xi > 0.5$, a predicted win was coded.

If $\pi xi < 0.5$, 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 second round models were used to predict the winners of the second round. This process continued for the third and higher rounds.

In 2014, a continuous process was used in verifying the models instead of doing round by round predictions as in 2013. Namely, a complete bracket was filled out in 2014 before any game was played. Results are given in Chapter 4.

CHAPTER 4. RESULTS

4.1. Development Models

4.1.1. Development of Least Squares Regression Model for the First Round

A least squares regression model to help predict the winning team for each game in the first round was developed and found to be:

$\hat{y}$=1.12250(Diff. in 3-pt goals)-0.44657(Diff. in free throws) +2.29479(Diff. in blocks) -1.68434(Diff. in Seeds)

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.1. Table 4.2 gives the steps associated with the stepwise selection technique and the associated R-square values as variables are added to the model. The model with all 4 variables explains an estimated 76% of the variation in point spread.

Table 4.1. Point Spread Model Parameter Estimates

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
3-pt Goal	1.12250	0.38797	965.36546	8.37	0.0053
Free Throws	-0.44657	0.21710	487.94901	4.23	0.0440
Blocks	2.29479	0.86548	810.75238	7.03	0.0102
Seed	-1.68434	0.17355	10863	94.19	<.0001

Table 4.2. Summary of Stepwise Selection for Point Spread Model

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	F Value	Pr > F
1	Seed		1	0.6935	0.6935	142.58	<.0001
2	Blocks		2	0.0307	0.7243	6.91	0.0108
3	3-pt Goal		3	0.0208	0.7451	4.99	0.0292
4	Free Throws		4	0.0168	0.7619	4.23	0.0440

4.1.2. 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:

$$\pi\,(DIS, DFG) = \frac{e^{0.279*DFG-0.418*DIS}}{1+ e^{0.279*DFG-0.418*DIS}}$$

Where π (DIS, DFG) is the estimated probability that the team of interest will win the game with DIS and DFG in model.

Table 4.3 shows the steps for the stepwise selection technique and Table 4.4 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 4.5 shows the Hosmer and Lemeshow test was done to test whether there was evidence the logistic model was not appropriate. The p-value was 0.907 indicating that there was no evidence to reject using the logistic model.

Table 4.3. Summary of Stepwise Selection for Logistic Regression Model

Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	DIS		1	1	39.1351		<.0001
2	DFG		1	2	4.9125		0.0267

Table 4.4. Logistic Regression Model Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
DFG	1	0.2790	0.1385	4.0616	0.0439
DIS	1	-0.4180	0.1212	11.9038	0.0006

Table 4.5. Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
4.0662	9	0.9070

4.1.3. Development of Least Squares Regression Model for the Second Round

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

$\hat{y}$ =1.34571(Diff. in Field Goals) + 0.54848(Diff. in Average Points)

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.6. Table 4.7 gives the steps associated with the stepwise selection technique and the associated R-square values as variables are added to the model. The model with all 2 variables explains an estimated 56% of the variation in point spread.

Table 4.6. Point Spread Model Parameter Estimates

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Field Goal	1.34571	0.75138	339.26527	3.21	0.0834
Average Points	0.54848	0.35356	254.54023	2.41	0.1313

Table 4.7. 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	Field Goal		1	0.5230	0.5230	-1.3703	33.99	<.0001
2	Average Points		2	0.0354	0.5584	-1.4964	2.41	0.1313

4.1.4. 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:

$$\pi\,(\text{DFG}) = \frac{e^{0.3538*DFG}}{1+ e^{0.3538*DFG}}$$

Where π (DFG) is the estimated probability that the team of interest will win the game with DFG in model.

Table 4.8 shows the steps for the stepwise selection technique and Table 4.9 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 4.10 shows the Hosmer and Lemeshow test was done to test whether there was evidence the logistic model was not appropriate. The p-value was 0.3354 indicating that there was no evidence to reject using the logistic model.

Table 4.8. Summary of Stepwise Selection for Logistic Regression Model

Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	DFG		1	1	10.4074		0.0013

Table 4.9. Logistic Regression Model Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
DFG	1	0.3538	0.1306	7.3374	0.0068

Table 4.10. Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
10.1889	9	0.3354

4.1.5. Development of Least Squares Regression Model for the Third and Higher Rounds

A least squares regression model to help predict the winning team for each game in the third and higher rounds was developed and found to be:

$\hat{y}$= 2.52646(Diff. in Assists) +1.18735(Diff. in Steals)-2.89252(Diff. in Seeds)

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 4.11. Table 4.12 gives the steps associated with the stepwise selection technique and the associated R-square values as variables are added to the model. The model with all 3 variables explains an estimated 68% of the variation in point spread.

Table 4.11. Point Spread Model Parameter Estimates

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
assists	2.52646	0.60519	2309.17242	17.43	0.0003
steals	1.18735	0.63843	458.29319	3.46	0.0738
seed	-2.89252	0.56332	3493.47910	26.37	<.0001

Table 4.12. 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	average points		1	0.3956	0.3956	20.5185	18.99	0.0002
2	seed		2	0.1819	0.5776	7.9149	12.06	0.0017
3	assists		3	0.0718	0.6493	4.1519	5.53	0.0263
4		average points	2	0.0136	0.6357	3.2441	1.05	0.3151
5	steals		3	0.0414	0.6771	1.9232	3.46	0.0738

4.1.6. 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

$$\pi\,(\text{DAP, DIS}) = \frac{e^{0.3222*DAP-0.5494*DIS}}{1+e^{0.3222*DAP-0.5494*DIS}}$$

Where π (DAP, DIS) is the estimated probability that the team of interest will win the game with DAP and DIS in model.

Table 4.13 shows the steps for the stepwise selection technique and Table 4.14 gives the parameter estimates, their standard errors and associated p-values when all the variables are in the model. Table 4.15 shows the Hosmer and Lemeshow test was done to test whether there was evidence the logistic model was not appropriate. The p-value was 0.0811 indicating that there was no evidence to reject using the logistic model.

Table 4.13. Summary of Stepwise Selection for Logistic Regression Model

Step	Effect		DF	Number	Score	Wald	Pr > ChiSq
	Entered	Removed		In	Chi-Square	Chi-Square	
1	DAP		1	1	10.1130		0.0015
2	DIS		1	2	8.1841		0.0042
3	DI3G		1	3	3.0999		0.0783
4	DIAS		1	4	4.0119		0.0452
5		DAP	1	3		1.7780	0.1824
6		DI3G	1	2		0.8043	0.3698
7		DIAS	1	1		1.3890	0.2386
8	DAP		1	2	8.1760		0.0042

Table 4.14. Logistic Regression Model Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
DAP	1	0.3222	0.1665	3.7472	0.0529
DIS	1	-0.5494	0.3062	3.2182	0.0728

Table 4.15. Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
14.0254	8	0.0811

4.2. Prediction Round by Round Using Models Developed

The three least squares regression models were used to predict the first round, second round and third round through final of 2013 season to check the prediction accuracy of the models. It is noted that the 2013 season was not used in the development of the models.

Tables 4.16 – 4.18 give the results as to how accurately the least squares regression models for each of the rounds of the NCAA 2013 women's basketball tournament with the results of the third through the sixth rounds combined together.

Tables 4.19 – 4.21 give similar results for the logistic models.

Table 4.16. Accuracy of Least Squares Regression Model When Predicting First Round of 2013

Point Spread		**Predicted**		
		Win	Loss	Total
Actual	Win	13	3	16
	Loss	1	15	16
	Total	14	18	32
Overall Accuracy				87.5%

Table 4.17. Accuracy of Least Squares Regression Model When Predicting Second Round of 2013

Point Spread		**Predicted**		
		Win	Loss	Total
Actual	Win	7	1	8
	Loss	2	6	8
	Total	9	7	16
Overall Accuracy				81.3%

Table 4.18. Accuracy of Least Squares Regression Model When Predicting Third and Higher Rounds of 2013

Point Spread		**Predicted**		
		Win	Loss	Total
Actual	Win	3	2	5
	Loss	2	8	10
	Total	5	10	15
Overall Accuracy				73.3%

Table 4.19. Accuracy of Logistic Regression Model When Predicting First Round of 2013

Logistic		Predicted		
		Win	Loss	Total
Actual	Win	15	1	16
	Loss	2	14	16
	Total	17	15	32
Overall Accuracy				90.63%

Table 4.20. Accuracy of Logistic Regression Model When Predicting Second Round of 2013

Logistic		Predicted		
		Win	Loss	Total
Actual	Win	7	1	8
	Loss	2	6	8
	Total	9	7	16
Overall Accuracy				81.25%

Table 4.21. Accuracy of Logistic Regression Model When Predicting Third and Higher Rounds of 2013

Logistic		Predicted		
		Win	Loss	Total
Actual	Win	3	2	5
	Loss	2	8	10
	Total	5	10	15
Overall Accuracy				73.33%

It is noted that the percentage of accuracy for each round using the least squares models and using the logistic models are very close.

4.3. Bracketing the 2014 Tournament Before Tournament Begins

The accuracy of the least squares regression models were checked against each rounds of the games from 2014 season. For each round of the game, statistics for the significant factors were collected and the difference was taken and placed into the predictive models to find a predicted point spread, $\hat{y}$.

If $\hat{y} > 0$, a predicted win for the point spread model was coded.

If $\hat{y} < 0$, a predicted loss for the point spread model was coded.

Results were predicted for every round before the tournament begin. Variables associated with teams predicted to win the first round were placed into the second round model. Variables associated with team predicted to win the second round were placed in the third round model to predict which teams would win this round. This process continued.

These predicted results were then compared against the actual results for each round of the game for 2014.

4.4. Examples for Each Round of 2014 Tournament

4.4.1. Least Squares Regression Model for First Round

The following gives the least squares model for the first round:

$\hat{y}$=1.12250(Diff. in 3-pt goals)-0.44657(Diff. in free throws) +2.29479(Diff. in blocks)-1.68434(Diff. in Seeds)

Table 4.22. Michigan St. and Hampton Statistics

Team	Score	3-pt goals*	Free throws*	Blocks*	Seed
Michigan St.	91	34.276	71.164	4.613	5
Hampton	61	30.213	61.93	4.733	12
Difference	30	4.063	9.234	-0.12	-7

* Average per game for season

Using the model above, Michigan St. had a predicted point spread of:

$\hat{y}$ = 1.12250*4.063 - 0.44657*9.234 + 2.29479*(-0.12) - 1.68434*(-7) = 11.95

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Michigan St., who won the game by a score of 91-61.

Table 4.23. South Carolina and Cal St. Northridge Statistics

Team	Score	3-pt goals*	Free throws*	Blocks*	Seed
South Carolina	73	34.551	66.826	7.258	1
Cal St. Northridge	58	32.725	68.859	4	16
Difference	15	1.826	-2.033	3.258	-15

* Average per game for season

Using the model above, South Carolina had a predicted point spread of:

$\hat{y}$ = 1.12250*1.826 - 0.44657*(-2.033) + 2.29479*3.258 - 1.68434*(-15) = 35.70

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for South Carolina, who won the game by a score of 73-58.

Table 4.24. Middle Tenn. and Oregon St. Statistics

Team	Score	3-pt goals*	Free throws*	Blocks*	Seed
Middle Tenn.	36	28.135	64	1.839	8
Oregon St.	55	37.068	65.802	6.182	9
Difference	-19	-8.933	-1.802	-4.343	-1

* Average per game for season

Using the model above, Middle Tenn. had a predicted point spread of:

$\hat{y}$ = 1.12250*(-8.933) - 0.44657*(-1.802) + 2.29479*(-4.343) - 1.68434*(-1) = -17.50

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Middle Tenn., who lost the game by a score of 36-55.

Table 4.25. North Carolina and UT Martin Statistics

Team	Score	3-pt goals*	Free throws*	Blocks*	Seed
North Carolina	60	32.573	66.709	4.758	4
UT Martin	58	35.959	74.934	2.677	13
Difference	2	-3.386	-8.225	2.081	-9

* Average per game for season

Using the model above, North Carolina had a predicted point spread of:

$\hat{y} = 1.12250*(-3.386) - 0.44657*(-8.225) + 2.29479*2.081 - 1.68434*(-9) = 19.81$

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for North Carolina, who won the game by a score of 60-58.

Table 4.26. Western Ky. and Baylor Statistics

Team	Score	3-pt goals*	Free throws*	Blocks*	Seed
Western Ky.	74	32.981	72.904	3.467	15
Baylor	87	33.458	72.761	4.212	2
Difference	-13	-0.477	0.143	-0.745	13

* Average per game for season

Using the model above, Western Ky. had a predicted point spread of:

$\hat{y} = 1.12250*(-0.477) - 0.44657*0.143 + 2.29479*(-0.745) - 1.68434*(13) = -24.21$

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Western Ky., who lost the game by a score of 74-87.

Table 4.27. Chattanooga and Syracuse Statistics

Team	Score	3-pt goals*	Free throws*	Blocks*	Seed
Chattanooga	53	31.81	72.727	4.813	11
Syracuse	59	31.874	72.804	4.387	6
Difference	-6	-0.064	-0.077	0.426	5

* Average per game for season

Using the model above, Chattanooga had a predicted point spread of:

$\hat{y}$ = 1.12250*(-0.064) - 0.44657*(-0.077) + 2.29479*0.426 - 1.68434*(5) = -7.48

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for Chattanooga, who lost the game by a score of 53-59.

Table 4.28. Robert Morris and Notre Dame Statistics

Team	Score	3-pt goals*	Free throws*	Blocks*	Seed
Robert Morris	42	33.929	68.829	3.387	16
Notre Dame	93	40.648	75.217	4.094	1
Difference	-51	-6.719	-6.388	-0.707	15

* Average per game for season

Using the model above, Robert Morris had a predicted point spread of:

$\hat{y}$ = 1.12250*(-6.719) - 0.44657*(-6.388) + 2.29479*(-0.707) - 1.68434*(15) = -31.58

Since $\hat{y}$ <0 this game was coded as a correctly predicted loss for Robert Morris, who lost the game by a score of 42-93.

Table 4.29. Albany (NY) and West Virginia Statistics

Team	Score	3-pt goals*	Free throws*	Blocks*	Seed
Albany (NY)	61	30.678	67.76	2.375	15
West Virginia	76	33.109	69.415	4.455	2
Difference	-15	-2.431	-1.655	-2.08	13

* Average per game for season

Using the model above, Albany (NY) had a predicted point spread of:

$\hat{y}$ = 1.12250*(-2.431) - 0.44657*(-1.655) + 2.29479*(-2.08) - 1.68434*(13) = -28.66

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Albany (NY), who lost the game by a score of 61-76.

Round 1:

Number correct: 26

Number incorrect: 6

Total: 32

4.4.2. Least Squares Regression Model for Second Round:

The following gives the least squares model for the second round:

$\hat{y}$ = 1.34571(Diff. in Field Goals) + 0.54848(Diff. in Average Points)

Table 4.30. South Carolina and Oregon St. Statistics

Team	Score	Field Goals*	Average Points*
South Carolina	78	48.238	73.1935484
Oregon St.	69	43.987	70.969697
Difference	9	4.251	2.2238514

* Average per game for season

Using the model above, South Carolina had a predicted point spread of:

$\hat{y}$ = 1.34571*4.251 + 0.54848*2.2238514= 6.94

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for South Carolina, who won the game by a score of 78-69.

Table 4.31. DePaul and Duke Statistics

Team	Score	Field Goals*	Average Points*
DePaul	65	45.045	83.7272727
Duke	74	49.876	80.2424242
Difference	9	-4.831	3.4848485

* Average per game for season

Using the model above, DePaul had a predicted point spread of:

$\hat{y}$ = 1.34571*(-4.831) + 0.54848*3.4848485 = -4.59

Since $\hat{y} < 0$ this game was coded as an incorrectly predicted loss for DePaul, who won the game by a score of 74-65.

Table 4.32. Maryland and Texas Statistics

Team	Score	Field Goals*	Average Points*
Maryland	69	48.956	83.1666667
Texas	64	43.499	69.5625
Difference	5	5.457	13.6041667

* Average per game for season

Using the model above, Maryland had a predicted point spread of:

$\hat{y}$ = 1.34571*5.457 + 0.54848*13.6041667 = 14.81

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for Maryland, who won the game by a score of 69-64.

Table 4.33. Kentucky and Syracuse Statistics

Team	Score	Field Goals*	Average Points*
Kentucky	64	43.478	81.34375
Syracuse	59	39.832	73.483871
Difference	5	3.646	7.859879

* Average per game for season

Using the model above, Kentucky had a predicted point spread of:

$\hat{y}$ = 1.34571*3.646 + 0.54848*7.859879 = 9.22

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Kentucky, who won the game by a score of 64-59.

Round 2:

Number correct: 9

Number incorrect: 7

Total: 16

4.4.3. Least Squares Regression Model for Third Round:

The following gives the least squares model for the third and higher rounds:

$\hat{y}$ = 2.52646(Diff. in Assists) +1.18735(Diff. in Steals)-2.89252(Diff. in Seeds)

Table 4.34. South Carolina and North Carolina Statistics

Team	Score	Assists*	Steals*	Seed
South Carolina	58	14.742	6.129	1
North Carolina	65	15.727	11.636	4
Difference	-7	-0.985	-5.507	-3

* Average per game for season

Using the model above, South Carolina had a predicted point spread of:

$\hat{y}$ = 2.52646*(-0.985) +1.18735*(-5.507)-2.89252*(-3) = -0.35

Since $\hat{y}$ < 0 this game was coded as a correctly predicted loss for South Carolina, who lost the game by a score of 58-65.

Table 4.35. Baylor and Kentucky Statistics

Team	Score	Assists*	Steals*	Seed
Baylor	90	18.697	7.212	2
Kentucky	72	14.219	9.969	3
Difference	18	4.478	-2.757	-1

* Average per game for season

Using the model above, Baylor had a predicted point spread of:

$\hat{y}$ = 2.52646*4.478 +1.18735*(-2.757)-2.89252*(-1) = 10.93

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Baylor, who won the game by a score of 90-72.

Round 3:

Number correct: 6

Number incorrect: 2

Total: 8

4.4.4. Least Squares Regression Model for Fourth Round:

The following gives the least squares model for the third and higher rounds:

$\hat{y}$ = 2.52646(Diff. in Assists) +1.18735(Diff. in Steals)-2.89252(Diff. in Seeds)

Table 4.36. North Carolina and Stanford Statistics

Team	Score	Assists*	Steals*	Seed
North Carolina	65	15.727	11.636	4
Stanford	74	17.813	5.781	2
Difference	-9	-2.086	5.855	2

* Average per game for season

Using the model above, North Carolina had a predicted point spread of:

$\hat{y}$ = 2.52646(Diff. in Assists) +1.18735(Diff. in Steals)-2.89252(Diff. in Seeds) = -4.10

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for North Carolina, who lost the game by a score of 65-74.

Table 4.37. Baylor and Notre Dame Statistics

Team	Score	Assists*	Steals*	Seed
Baylor	69	18.697	7.212	2
Notre Dame	88	20.688	9.625	1
Difference	-19	-1.991	-2.413	1

* Average per game for season

Using the model above, Baylor had a predicted point spread of:

$\hat{y}$ = 2.52646*(-1.991) +1.18735*(-2.413)-2.89252*1 = -10.79

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Baylor, who lost the game by a score of 69-88.

Round 4:

Number correct: 4

Number incorrect: 0

Total: 4

4.4.5. Least Squares Regression Model for Fifth Round

The following gives the least squares model for the third and higher rounds:

$\hat{y}$ = 2.52646(Diff. in Assists) +1.18735(Diff. in Steals)-2.89252(Diff. in Seeds)

Table 4.38. UConn and Stanford Statistics

Team	Score	Assists*	Steals*	Seed
UConn	75	21.559	9.765	1
Stanford	56	17.813	5.781	2
Difference	19	3.746	3.984	-1

* Average per game for season

Using the model above, UConn had a predicted point spread of:

$\hat{y}$ = 2.52646*3.746 +1.18735*3.984-2.89252*(-1) = 17.09

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for UConn, who won the game by a score of 75-56.

Table 4.39. Maryland and Notre Dame Statistics

Team	Score	Assists*	Steals*	Seed
Maryland	61	19.6	8.3	4
Notre Dame	87	20.688	9.625	1
Difference	-26	-1.088	-1.325	3

* Average per game for season

Using the model above, Maryland had a predicted point spread of:

$\hat{y}$ = 2.52646*(-1.088) +1.18735(-1.325)-2.89252*3 = -13.00

Since $\hat{y} < 0$ this game was coded as a correctly predicted loss for Maryland, who lost the game by a score of 61-87.

Round 5:

Number correct: 2

Number incorrect: 0

Total: 2

4.4.6. Least Squares Regression Model for Sixth Round

The following gives the least squares model for the third and higher rounds:

$\hat{y}$ = 2.52646(Diff. in Assists) +1.18735(Diff. in Steals)-2.89252(Diff. in Seeds)

Table 4.40. UConn and Notre Dame Statistics

Team	Score	Assists*	Steals*	Seed
UConn	79	21.559	9.765	1
Notre Dame	58	20.688	9.625	1
Difference	21	0.871	0.14	0

* Average per game for season

Using the model above, UConn had a predicted point spread of:

$\hat{y}$ = 2.52646*0.871 +1.18735*0.14-2.89252*0 = 2.37

Since $\hat{y} > 0$ this game was coded as a correctly predicted win for UConn, who won the game by a score of 79-58.

Round 6:

Number correct: 1

Number incorrect: 0

Total: 1

A summary of the number of correct and incorrect predictions for each round of the 2014 tournament is given in Table 4.42.

Least square regression models were used to predict each round of NCAA women's basketball tournament of 2013 and 2014.

Table 4.41. Prediction Results of Each Round for 2013: (Least Squares Regression Model)

	Correct	**Incorrect**	**Total games**
First round	28	4	32
Second round	13	3	16
Third round	6	2	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	0	1	1
Overall Accuracy			79.37%

Table 4.42. Prediction Results of Each Round for 2014: (Least Squares Regression Model)

	Correct	**Incorrect**	**Total games**
First round	26	6	32
Second round	9	7	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%

Figure 3 and Figure 4 show the predicted results of 2013 tournament and 2014 tournament when least squares regression models were used. The highlighted parts of both figures are the incorrectly predicted results.

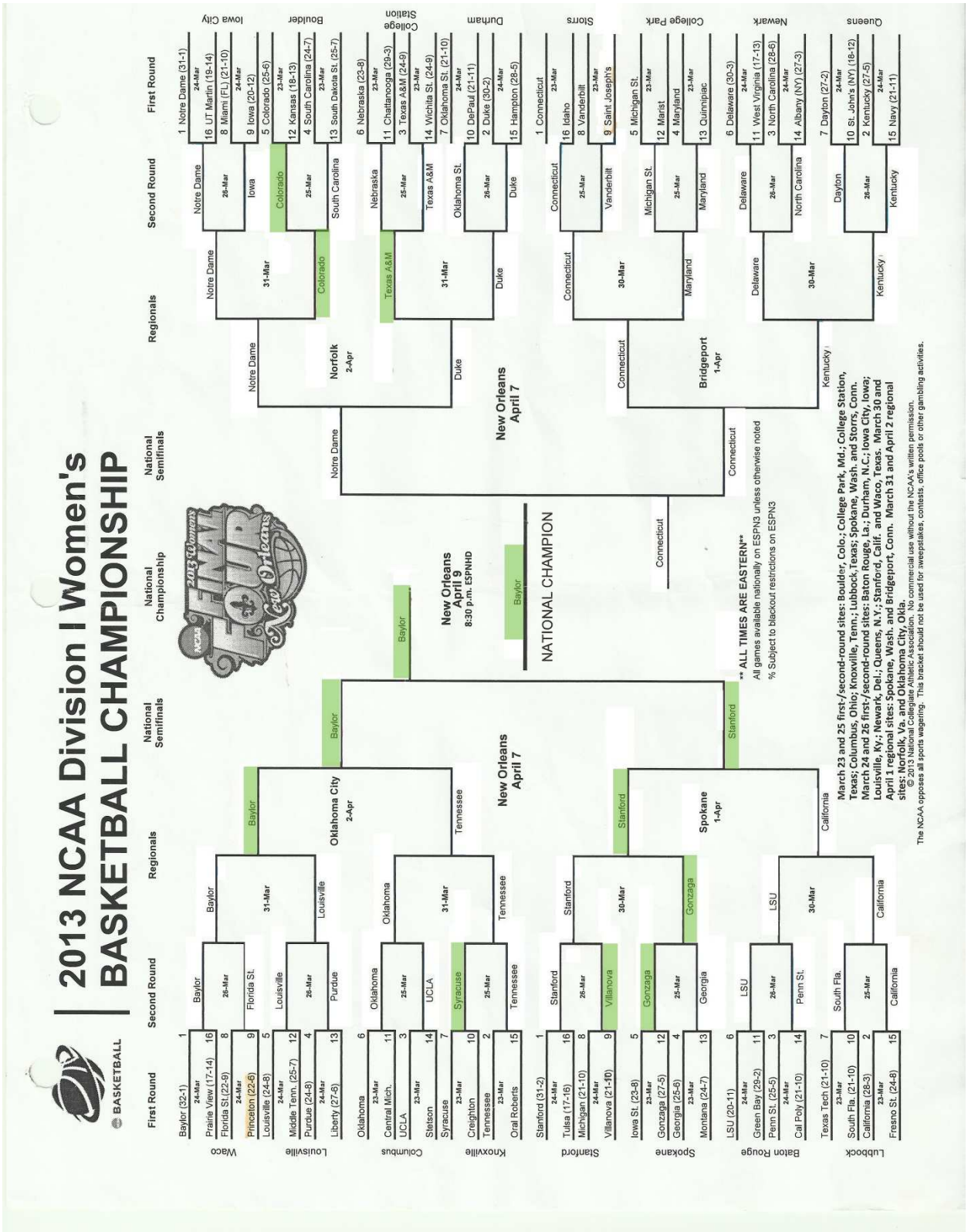

Figure 3: Prediction of the NCAA women's basketball tournament bracket for 2013 season

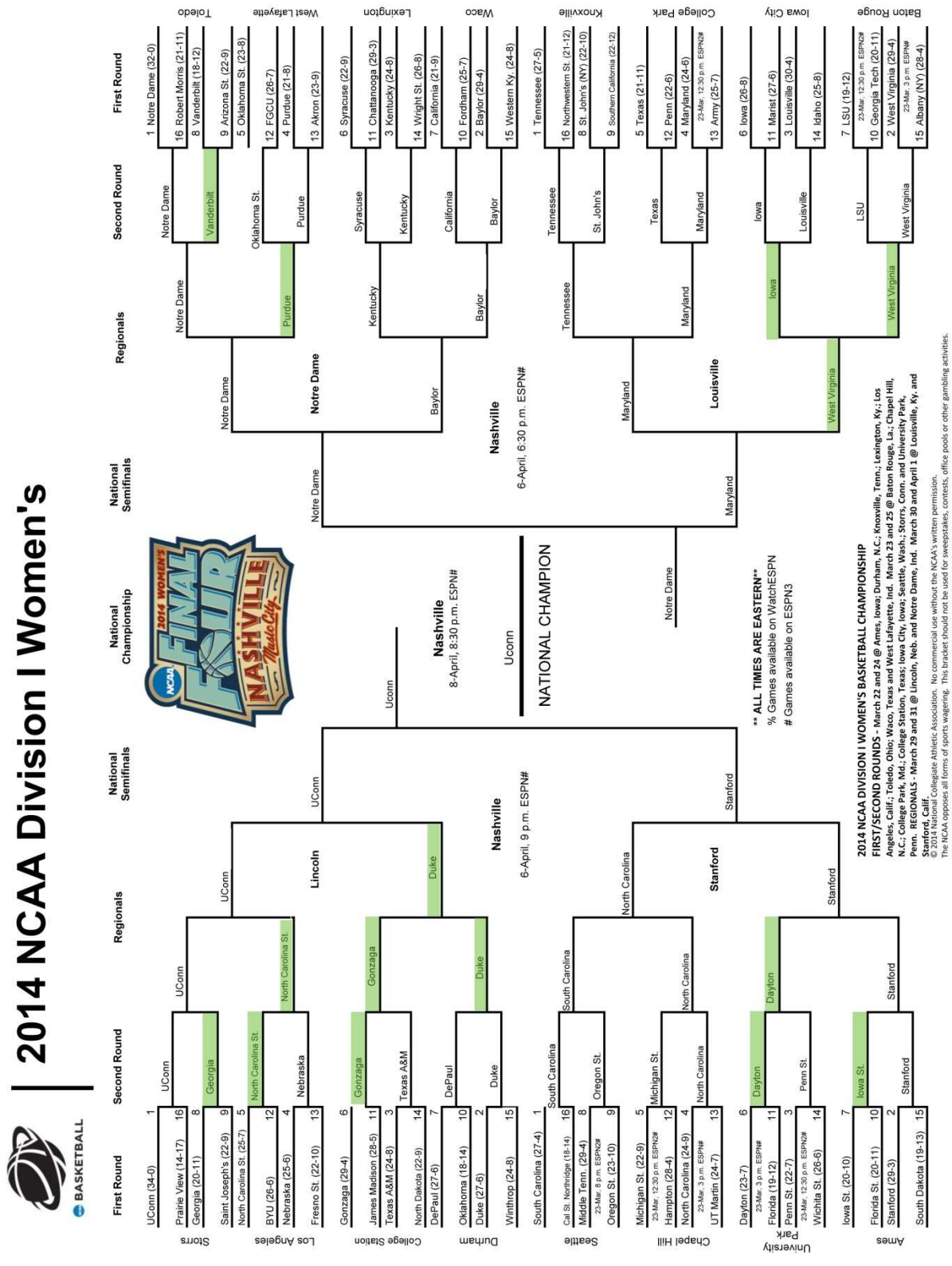

Figure 4: Prediction of the NCAA women's basketball tournament bracket for 2014 season

CHAPTER 5. CONCLUSION

To verify the accuracy of prediction results for the least squares regression model, differences of the seasonal averages for both teams for all previously mentioned variables were placed in the model developed for Round 1, Round 2 and Rounds 3-6. The least squares regression model and the logistic regression model for the first round had approximately a 87.5% and 90.6% chance of correctly predicting the results, respectively. The least squares regression model and the logistic regression model for the second round had approximately a 81.3% and 81.2% chance of correctly predicting the results, respectively. The least squares regression model and the logistic regression model for the third and higher round had approximately a 73.3% and 73.3% chance of correctly predicting the results, respectively.

In 2014, a continuous process was used in verifying the models instead of doing round by round predictions as in 2013. Namely, a complete bracket was filled out in 2014 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 least squares models, the models had approximately a 76% chance of correctly predicting the winner of a basketball game.

REFERENCES

Abraham, B., and J. Ledolter. 2006. Introduction to Regression Modeling, (1 st ed.). Belmont CA: Thomson Brooks/Cole.

Carlin, B.P. 1996. Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information. The American Statistician 50:39-43.

Caudill, S.B. 2003. Predicting Discrete Outcomes with the Maximum Score Estimator: the Case of the NCAA Men's Basketball Tournament. International Journal of Forecasting 19:313-317.

Duncan, M.C. 2006. Gender warriors in sport: Women and the media, In A. A. Raney & J. Bryant (Eds.), Handbook of sports and media (pp. 231-252). Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M.J. 1996. Media coverage of the post Title IX female athlete: A feminist analysis of sport, gender, and power. Duke Journal of Gender Law & Public Policy 3(1):95-127.

Kian, E.T.M., M. Mondello, and J. Vincent. 2009. ESPN- The Women's Sports Network? A Content Analysis of Internet Coverage of March Madness. Journal of Broadcasting & Electronic Media 53:477-495.

Kubatko, J., D. Oliver, K. Pelton, and D.T. Rosenbaum. 2007. A Starting Point for Analyzing Basketball Statistics. Journal of Quantitative Analysis in Sports 53:94-98.

Magel, R., and S. Unruh. 2013. Determining Factors Influencing the Outcome of College Basketball Games. Open Journal of Statistics (http://www.scirp.org/journal/ojs).

Schwertman, N.C., K.L. Schenk, and B.C. Holbrook. 1993. More Probability Models for the NCAA Regional Basketball Tournaments. The American Statistician, 50:34-38.

Smith, T., and N.C. Schwertman. 1999. Can the NCAA Basketball Tournament Seeding be Used to Predict Margin of Victory?. The American Statistician. 53:94-98.

West, B.T. 2006. A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament", Journal of Quantitative Analysis in Sports, 2(3):3-8.

Zhang, X. 2013. Bracketing NCAA Men's Division I Basketball Tournament. Unpublished Master's Thesis Paper, Department of Statistics, North Dakota State University.

APPENDIX. SAS CODE

Code for least squares regression model for first round

```
/* ---------------------------------------------------------------------
   Code generated by SAS Task

   Generated on: Sunday, April 27, 2014 at 2:17:32 PM
   By task: Linear Regression2

   Input Data: WORK.DATA
   Server:  Local
   --------------------------------------------------------------------- */
ODS GRAPHICS ON;

%_eg_conditional_dropds(WORK.SORTTempTableSorted,
                WORK.TMP1TempTableForPlots);
/* ---------------------------------------------------------------------
   Determine the data set's type attribute (if one is defined)
   and prepare it for addition to the data set/view which is
   generated in the following step.
   --------------------------------------------------------------------- */
DATA _NULL_;
        dsid = OPEN("WORK.DATA", "I");
        dstype = ATTRC(DSID, "TYPE");
        IF TRIM(dstype) = " " THEN
                DO;
                CALL SYMPUT("_EG_DSTYPE_", "");
                CALL SYMPUT("_DSTYPE_VARS_", "");
                END;
        ELSE
                DO;
                CALL SYMPUT("_EG_DSTYPE_", "(TYPE=""" || TRIM(dstype) || """)");
                IF VARNUM(dsid, "_NAME_") NE 0 AND VARNUM(dsid, "_TYPE_") NE 0 THEN
                        CALL SYMPUT("_DSTYPE_VARS_", "_TYPE_ _NAME_");
                ELSE IF VARNUM(dsid, "_TYPE_") NE 0 THEN
                        CALL SYMPUT("_DSTYPE_VARS_", "_TYPE_");
                ELSE IF VARNUM(dsid, "_NAME_") NE 0 THEN
                        CALL SYMPUT("_DSTYPE_VARS_", "_NAME_");
                ELSE
                        CALL SYMPUT("_DSTYPE_VARS_", "");
                END;
        rc = CLOSE(dsid);
        STOP;
RUN;

/* ---------------------------------------------------------------------
   Data set WORK.DATA does not need to be sorted.
   --------------------------------------------------------------------- */
DATA WORK.SORTTempTableSorted &_EG_DSTYPE_ / VIEW=WORK.SORTTempTableSorted;
```

```
        SET WORK.DATA(KEEP=pointspread DIFG DI3G DIFT DIAR DIAA DIAB DIAS DIAP DIS
&_DSTYPE_VARS_);
RUN;
TITLE;
TITLE1 "Linear Regression Results";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";
PROC REG DATA=WORK.SORTTempTableSorted
                PLOTS(ONLY)=ALL
        ;
        Linear_Regression_Model: MODEL pointspread = DIFG DI3G DIFT DIAR DIAA DIAB DIAS
DIAP DIS
                /               SELECTION=STEPWISE
                SLE=0.15
                SLS=0.15
                INCLUDE=0
                NOINT
        ;
RUN;
QUIT;

/* ---------------------------------------------------------------
   End of task code.
   --------------------------------------------------------------- */
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORTTempTableSorted,
                WORK.TMP1TempTableForPlots);
TITLE; FOOTNOTE;
        ODS GRAPHICS OFF;
```

Code for least squares regression model for second round

```
/* ---------------------------------------------------------------
   Code generated by SAS Task

   Generated on: Sunday, April 27, 2014 at 2:23:00 PM
   By task: Linear Regression3

   Input Data: WORK.DATA
   Server:  Local
   --------------------------------------------------------------- */
ODS GRAPHICS ON;

%_eg_conditional_dropds(WORK.SORTTempTableSorted,
                WORK.TMP1TempTableForPlots);
/* ---------------------------------------------------------------
   Determine the data set's type attribute (if one is defined)
```

```
  and prepare it for addition to the data set/view which is
  generated in the following step.
  ------------------------------------------------------------------- */
DATA _NULL_;
        dsid = OPEN("WORK.DATA", "I");
        dstype = ATTRC(DSID, "TYPE");
        IF TRIM(dstype) = " " THEN
                DO;
                CALL SYMPUT("_EG_DSTYPE_", "");
                CALL SYMPUT("_DSTYPE_VARS_", "");
                END;
        ELSE
                DO;
                CALL SYMPUT("_EG_DSTYPE_", "(TYPE=""" || TRIM(dstype) || """)");
                IF VARNUM(dsid, "_NAME_") NE 0 AND VARNUM(dsid, "_TYPE_") NE 0 THEN
                        CALL SYMPUT("_DSTYPE_VARS_", "_TYPE_ _NAME_");
                ELSE IF VARNUM(dsid, "_TYPE_") NE 0 THEN
                        CALL SYMPUT("_DSTYPE_VARS_", "_TYPE_");
                ELSE IF VARNUM(dsid, "_NAME_") NE 0 THEN
                        CALL SYMPUT("_DSTYPE_VARS_", "_NAME_");
                ELSE
                        CALL SYMPUT("_DSTYPE_VARS_", "");
                END;
        rc = CLOSE(dsid);
        STOP;
RUN;

/* ------------------------------------------------------------------
  Data set WORK.DATA does not need to be sorted.
  ------------------------------------------------------------------- */
DATA WORK.SORTTempTableSorted &_EG_DSTYPE_ / VIEW=WORK.SORTTempTableSorted;
        SET WORK.DATA(KEEP=pointspread DIFG DI3G DIFT DIAR DIAA DIAB DIAS DIAP DIS
&_DSTYPE_VARS_);
RUN;
TITLE;
TITLE1 "Linear Regression Results";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";
PROC REG DATA=WORK.SORTTempTableSorted
                PLOTS(ONLY)=ALL
        ;
        Linear_Regression_Model: MODEL pointspread = DIFG DI3G DIFT DIAR DIAA DIAB DIAS
DIAP DIS
                /               SELECTION=STEPWISE
                SLE=0.15
                SLS=0.15
                INCLUDE=0
                NOINT
        ;
```

```
RUN;
QUIT;

/* ---------------------------------------------------------------
   End of task code.
   --------------------------------------------------------------- */
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORTTempTableSorted,
                WORK.TMP1TempTableForPlots);
TITLE; FOOTNOTE;
        ODS GRAPHICS OFF;
```

Code for least squares regression model for third and higher rounds

```
/* ---------------------------------------------------------------
   Code generated by SAS Task

   Generated on: Sunday, April 27, 2014 at 2:26:40 PM
   By task: Linear Regression4

   Input Data: WORK.DATA
   Server:  Local
   --------------------------------------------------------------- */
ODS GRAPHICS ON;

%_eg_conditional_dropds(WORK.SORTTempTableSorted,
                WORK.TMP1TempTableForPlots);
/* ---------------------------------------------------------------
   Determine the data set's type attribute (if one is defined)
   and prepare it for addition to the data set/view which is
   generated in the following step.
   --------------------------------------------------------------- */
DATA _NULL_;
        dsid = OPEN("WORK.DATA", "I");
        dstype = ATTRC(DSID, "TYPE");
        IF TRIM(dstype) = " " THEN
                DO;
                CALL SYMPUT("_EG_DSTYPE_", "");
                CALL SYMPUT("_DSTYPE_VARS_", "");
                END;
        ELSE
                DO;
                CALL SYMPUT("_EG_DSTYPE_", "(TYPE=""" || TRIM(dstype) || """)");
                IF VARNUM(dsid, "_NAME_") NE 0 AND VARNUM(dsid, "_TYPE_") NE 0 THEN
                        CALL SYMPUT("_DSTYPE_VARS_", "_TYPE_ _NAME_");
                ELSE IF VARNUM(dsid, "_TYPE_") NE 0 THEN
                        CALL SYMPUT("_DSTYPE_VARS_", "_TYPE_");
                ELSE IF VARNUM(dsid, "_NAME_") NE 0 THEN
                        CALL SYMPUT("_DSTYPE_VARS_", "_NAME_");
```

```
                ELSE
                        CALL SYMPUT("_DSTYPE_VARS_", "");
                END;
        rc = CLOSE(dsid);
        STOP;
RUN;

/* ---------------------------------------------------------------
   Data set WORK.DATA does not need to be sorted.
   --------------------------------------------------------------- */
DATA WORK.SORTTempTableSorted &_EG_DSTYPE_ / VIEW=WORK.SORTTempTableSorted;
        SET WORK.DATA(KEEP=pointspread DIFG DI3G DIFT DIAR DIAA DIAB DIAS DIAP DIS
&_DSTYPE_VARS_);
RUN;
TITLE;
TITLE1 "Linear Regression Results";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";
PROC REG DATA=WORK.SORTTempTableSorted
                PLOTS(ONLY)=ALL
        ;
        Linear_Regression_Model: MODEL pointspread = DIFG DI3G DIFT DIAR DIAA DIAB DIAS
DIAP DIS
                /               SELECTION=STEPWISE
                SLE=0.15
                SLS=0.15
                INCLUDE=0
                NOINT
        ;
RUN;
QUIT;

/* ---------------------------------------------------------------
   End of task code.
   --------------------------------------------------------------- */
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORTTempTableSorted,
                WORK.TMP1TempTableForPlots);
TITLE; FOOTNOTE;
ODS GRAPHICS OFF;
```

Code for logistic regression model for first round

```
/* ---------------------------------------------------------------
   Code generated by SAS Task
```

```
  Generated on: Sunday, April 27, 2014 at 2:31:45 PM
  By task: Logistic Regression

  Input Data: WORK.DATA
  Server:  Local
  ---------------------------------------------------------------- */
ODS GRAPHICS ON;

%_eg_conditional_dropds(WORK.SORTTempTableSorted);
/* ----------------------------------------------------------------
  Sort data set WORK.DATA
  ---------------------------------------------------------------- */

PROC SQL;
        CREATE VIEW WORK.SORTTempTableSorted AS
                SELECT T.pointspread, T.DIFG, T.DI3G, T.DIFT, T.DIAR, T.DIAA, T.DIAB, T.DIAS,
T.DIAP, T.DIS
        FROM WORK.DATA as T
;
QUIT;
TITLE;
TITLE1 "Logistic Regression Results for first round";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";
PROC LOGISTIC DATA=WORK.SORTTempTableSorted
                PLOTS(ONLY)=ALL
        ;
        MODEL pointspread (Event = '1')=DIFG DI3G DIFT DIAR DIAA DIAB DIAS DIAP DIS
        /
                SELECTION=STEPWISE
                SLE=0.15
                SLS=0.15
                INCLUDE=0
                NOINT
                LACKFIT
                LINK=LOGIT
        ;
RUN;
QUIT;

/* ----------------------------------------------------------------
  End of task code.
  ---------------------------------------------------------------- */
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORTTempTableSorted);
TITLE; FOOTNOTE;
        ODS GRAPHICS OFF;
```

Code for logistic regression model for second round

```
/* ---------------------------------------------------------------
   Code generated by SAS Task

   Generated on: Sunday, April 27, 2014 at 2:37:00 PM
   By task: Logistic Regression 2

   Input Data: WORK.DATA
   Server:  Local
   --------------------------------------------------------------- */
ODS GRAPHICS ON;

%_eg_conditional_dropds(WORK.SORTTempTableSorted);
/* ---------------------------------------------------------------
   Sort data set WORK.DATA
   --------------------------------------------------------------- */

PROC SQL;
        CREATE VIEW WORK.SORTTempTableSorted AS
                SELECT T.pointspread, T.DIFG, T.DI3G, T.DIFT, T.DIAR, T.DIAA, T.DIAB, T.DIAS,
T.DIAP, T.DIS
        FROM WORK.DATA as T
;
QUIT;
TITLE;
TITLE1 "Logistic Regression Results for second round";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";
PROC LOGISTIC DATA=WORK.SORTTempTableSorted
                PLOTS(ONLY)=ALL
        ;
        MODEL pointspread (Event = '1')=DIFG DI3G DIFT DIAR DIAA DIAB DIAS DIAP DIS
        /
                SELECTION=STEPWISE
                SLE=0.15
                SLS=0.15
                INCLUDE=0
                NOINT
                LACKFIT
                LINK=LOGIT
        ;
RUN;
QUIT;

/* ---------------------------------------------------------------
   End of task code.
   --------------------------------------------------------------- */
```

```
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORTTempTableSorted);
TITLE; FOOTNOTE;
ODS GRAPHICS OFF;
```

Code for logistic regression model for third and higher rounds

```
/* ---------------------------------------------------------------
   Code generated by SAS Task

   Generated on: Sunday, April 27, 2014 at 2:44:33 PM
   By task: Logistic Regression 4

   Input Data: WORK.DATA
   Server:  Local
   --------------------------------------------------------------- */
ODS GRAPHICS ON;

%_eg_conditional_dropds(WORK.SORTTempTableSorted);
/* ---------------------------------------------------------------
   Sort data set WORK.DATA
   --------------------------------------------------------------- */

PROC SQL;
        CREATE VIEW WORK.SORTTempTableSorted AS
                SELECT T.pointspread, T.DIFG, T.DI3G, T.DIFT, T.DIAR, T.DIAA, T.DIAB, T.DIAS,
T.DIAP, T.DIS
        FROM WORK.DATA as T
;
QUIT;
TITLE;
TITLE1 "Logistic Regression Results for third and higher rounds";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";
PROC LOGISTIC DATA=WORK.SORTTempTableSorted
                PLOTS(ONLY)=ALL
        ;
        MODEL pointspread (Event = '1')=DIFG DI3G DIFT DIAR DIAA DIAB DIAS DIAP DIS
        /
                SELECTION=STEPWISE
                SLE=0.15
                SLS=0.15
                INCLUDE=0
                NOINT
                LACKFIT
                LINK=LOGIT
```

```
        ;
RUN;
QUIT;

/* ---------------------------------------------------------------
   End of task code.
   --------------------------------------------------------------- */
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORTTempTableSorted);
TITLE; FOOTNOTE;
ODS GRAPHICS OFF;
```