FORECASTING POINT SPREAD FOR WOMEN'S VOLLEYBALL

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

By

Deling Zhang

In Partial Fulfillment of the Requirements.
for the Degree of
MASTER OF SCIENCE

Major Department:
Applied Statistics

November 2016

Fargo, North Dakota

North Dakota State University
Graduate School

Title

FORECASTING POINT SPREAD FOR WOMEN'S VOLLEYBALL

By

Deling Zhang

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

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Rhonda Magel
Chair

Dr. Ronald C. Degges

Dr. Donna Terbizan

Approved:

11/18/2016
Date

Dr.Rhonda Magel
Department Chair

ABSTRACT

Volleyball has become a well-known and competitive sport with physical and technical performances over the years. The game results are determined by some important factors such as players, and the team's skills to succeed in a championship. In this research, we propose to analyze volleyball data by using a multiple linear regression model and a logistic regression model. We develop a multiple regression model using in-game statistics that explain the point spread of a volleyball game. We also develop a logistic regression model that estimates the probability of a team winning the game based on the in-game statistics. Both of the models are validated and then the point spread model is used to predict the results of a volleyball game replacing the in-game statistics with the averages of the in-game statistics based on the past two previous matches of both teams. Results are given.

ACKNOWLEDGEMENTS

I would never have finished my master paper without guidelines of my committee members and support from my family. I would like to express my deepest gratitude to my advisor Dr. Rhonda for her excellent advice, caring, and warm heart. During my research, she provided me the most welcomed atmosphere and helped me to develop my background in statistics. Finally, I would like to thank Dr. Degges who was willing to participate at my defense committee for the completion of the master program.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE. INTRODUCTION

The game of volleyball was originally called “mintonette”, created by William G. Morgan in 1895, who was an instructor at the Young Men’s Christian Association (YMCA), in Holyoke, Massachusetts (NCAA (2014a)). Volleyball has become a well-known and popular sport that is played professionally, as well as in recreational leagues. Today, there are more than 46 million Americans who play volleyball and it ranks only behind soccer among all participation sports. The game of volleyball has brought many benefits to people in communities because it can be enjoyed anywhere a net is able to be setup. A typical volleyball game has six players on each side. Six rotational spots on the court are changed every time a particular side serves the ball. The aim is to deliver the ball over the net and ground it, or the ball touches on the ground of the opposing side, while preventing the ball from touching the ground on their own side (NCAA (2014a)).

Volleyball rules may change based on the location in which it is played. If the playing area is indoors, the court must be 19 meters by 9 meters. Indoor courts can also include an attached area designed by a line three meters back from the centerline. If playing beach volleyball, the sand court will be 16 meters by 8 meters. The playing space must be free from any obstructions to a recommended height of approximately 7 meters from the playing surfaces. The net height is measured from the center of the playing court with an appropriate measure device. The two ends of the net must be at the same height from the playing surfaces and it must not exceed the official height by more than two centimeters (NCAA (2014a)).

The rules appear to have changed over the years in volleyball but one thing that has remained constant is that a team may not exceed three contacts with the ball before it goes to the opposite team. The ideal sequence of contacts is usually a pass, a set, and a hit. The terminology

has changed over the years. These skills were traditionally called bump, set and spike. The volleyball game begins with a serve of the ball. Then players take turns rotating around the court so each player has a chance to "serve" the ball. The ball can be "served" when the server bumps the ball across the net with a fist, or throws the ball into the air and strikes with a hand or arm to bump it across the net into the opposing team's area. The members of the opposing team will attempt to "save" it from hitting the ground and knock it to one of their own players or hit it back over the net. Even though a team is allowed to hit the ball up to three times before hitting the ball over the net, no individual may hit or touch the ball twice in a row and the ball cannot be held, lifted, or carried. It should be noted that a block is not considered a hit. Play will continue until one team fails by allowing the ball to touch the ground in its own court or they correctly return it to the opposing court. A point is awarded to the serving team if the opposing team makes a mistake. If the serving team fails however, then the receiving team has control of the ball and becomes the serving team (NCAA (2014b)).

The scoring for the game has also changed over time. Initially, points could only be scored by the serving team, and games went until one of the teams reached 15 points and having at least two more points than the opposite team. If they were leading by less than two points, they continue playing until a 2-point lead is established. Now, however, volleyball has changed to rally scoring. Essentially, teams score points whenever the other team make a mistake, and a point is awarded on every serve. Matches may also be played now with a set of three games with each game going to 25 points. Just like it was previously, a team must win each game by two points. If the score is tied with even numbers, both teams have to continue playing the game until a 2-point lead is obtained. Otherwise, points keep accumulating until one team wins with a margin of victory of two points, even if the score is greater than 25 points (NCAA (2014b)).

For this research, we will develop a model that explains the point margin of a volleyball game based on the in-game statistics. A model will also be developed that estimates the probability of a team winning the game based on various in-game statistics. The model will be validated and then also used for prediction of future volleyball games.

CHAPTER TWO. LITERATURE REVIEW

There are many sports analyses that have been published in recent years. Commonly, a sports analysis will contain a regression model for the prediction of particular sports' scores. Examples of modeling developed for different sports including basketball, football, and hockey maybe found in Long and Magel (2013), Melynkov and Magel (2014), Roith and Magel (2014), Unruh and Magel (2016), Wang and Magel (2014). There have not been many publications for women's volleyball, however, due to the limited access of data readily available. Giatsis (2008) did a study pertaining to men's beach volleyball. This study considers the overall performance of a volleyball team depending on many factors related to the game. The purpose of this study was to explore the differences in playing characteristics between winning and losing teams in Federation Internationale de Volleyball (FIVB) Men's Beach Volleyball World Tour Tournament. There were 59 matches or 118 sets of the 1st 2003 FIVB men's Beach Volleyball in Rhodes, Greece. The important skills that were analyzed were serve, attack, block and dig. The statistical analysis methods used were independent t-tests comparing the differences in those skills between winning 2-0 and losing 2-1. Researchers also used a discriminant function analysis to determine which skills contributed significantly to winning in matches with 2-0 and 2-1 scores. According to results, it appears that opponents' attack errors was the most important factor contributing to a team winning. (Giatsis, 2008)

Generally, there are six categories of volleyball statistics that include attack, setting, serving, passing, defense, and blocking. By definition, attack means an attempt is recorded any time a player attempts to attack or hit strategically the ball into the opponent's court. There are three possible outcomes of an attack attempt which include kill, attack error and zero attack (ball

stays in play). This study will consider number of kills and number of attack errors by each team in addition to the variables hit percentage and side-out percentage for each team.

By definition, "a kill is awarded to a player any time an attack is not returnable by the opposition and is a direct cause of the opponent not returning the ball, or any time the attack leads directly to a blocking error by the opposition. When the player is awarded with a kill, the player is also awarded an attack attempt at the same time" (NCAA (2014a)).

By definition, "an attack error is charged to a player whenever an attacker makes a hitting error. For example, the player hits the ball out of bounds, or hits the ball into the net, that leads to a four hit violation, would account for attack errors. A "0 Attack" is any attack attempt that remains in play by the opposition" (NCAA (2014a)).

By definition, "Hitting percentage (PCT) is often used as a tool to evaluate the effectiveness of hitters throughout a given span. Normally, there is a formula available for PCT so we do need to gather three statistics that need to be tracked and recorded: attack attempts, attack errors, and attack kills. The percentage is determined by subtracting the total number of errors from the total number of kills and dividing that number by the total attack attempts" (NCAA (2014a)).

A side-out in volleyball occurs when the team that served the ball scores a point if the serve causes the ball to hit the ground in the other team's court or the opponent hits it out of bounds. It can also occur if the serving team hits the ball into the net or touches it more than three times. Under the side-out scoring system, the first team to achieve 25 points could win the game at the end. The side-out percentage is calculated by dividing the serve receive points by the number of serve receive attempts, times 100.

We considered these four in-game statistics because these are kept by several teams. Other in-game statistics do exist but are not available for the majority of women's volleyball games.

CHAPTER THREE. METHODS

3.1. Introduction for NCAA

The NCAA Women's Volleyball, Division I, refers to one of three championships in women's athletics contested by the NCAA. According to the NCAA website, Division I has over 294 schools, organized in 30 conferences within 8 districts (1997-1998). The 8 districts are divided into 4 regions (NCAA (2014b)).

3.2. Research Objective for the Study

The purpose of this study is to develop a model that explains the point spread of an NCAA Division I Women's Volleyball game based on various in-game statistics, and then to use this model to predict which team will win the volleyball game ahead of time. A volleyball match consists of 3 to 5 sets. The first team to win 3 sets wins the match.

The dependent variable in the model is the difference in scores between Team A and Team B; namely Score Team A – Score Team B. The independent variables considered for inclusion in the model are the following in-game statistics: the difference in the number of kills, the difference in the number of errors, the difference in the side-out percentages, and the difference in the hitting percentages (NCAA, 2013). The differences are in the order Team A – Team B. In addition to the in-game statistics, three indicator variables are considered to be included in the model. These new variables indicate the number of the set or game played in the match. A match is won if the team wins the most sets or games out of 5 sets or games. A match may consist of 3, 4, or 5 sets. The indicator variables for the sets were defined as;

$$Indicator1 = \begin{cases} 1, & if\ set\ is\ 2 \\ 0, & otherwise \end{cases} \qquad Indicator2 = \begin{cases} 1, & if\ set\ is\ 3 \\ 0, & otherwise \end{cases}$$

$$Indicator3 = \begin{cases} 1, & if\ set\ is\ 4\ or\ 5 \\ 0, & otherwise \end{cases}$$

3.3. Data Collection for the Study

Data were collected from a sample of matches for two seasons of NCAA Women's Division I volleyball games. The data included in-game statistics from 108 matches with 657 sets or games in the years of 2013 and 2014 from 18 universities in Division 1 NCAA women's volleyball. For each of the 18 universities, we collected data from 3 home matches and 3 away matches. We collected data on following 4 variables for each team playing in each game: number of kills, numbers of errors, hitting percentage (PCT), and side out percentage (Side-out). The number of kills varies from 12 to 19, the number of errors varies from 2 to 8, the hitting percentage varies from 16.7 to 40, and the side out percentage varies from 51 to 77. Differences for each of these variables is found between the two teams playing a game in the order Team A minus Team B.

Tables 3.1 and 3.2 are examples of the collected data that show the values for the in-game statistics for Teams A and B from each set in the match. We want to estimate the set score margin; namely (Team A minus Team B) point spread. Table 3.3 gives the differences of each of the in-game statistics.

Table 3.1

Box Score Team A

	Team A Score	**Kill**	**Errors**	**PCT**	**Side-out**
Set1	25	19	8	30.6	58
Set2	25	15	3	40	64
Set3	25	16	7	25.7	77

Table 3.2

Box Score Team B

	Team B Score	Kill	Errors	PCT	Side-out
Set1	23	13	4	22.5	54
Set2	20	14	8	16.7	51
Set3	21	12	2	31.2	66

The two tables are examples of the box scores for each team. The differences between the scores and each of the in-game statistics are given in table 3.3. The differences between the scores is the point spread.

Table 3.3

Differences between Teams A and B

	Point spread	Kill	Errors	PCT	Side-out
Set1	2	6	4	8.1	4
Set2	5	1	-5	23.3	13
Set3	4	4	5	2.4	11

The differences are given in the order Team A minus Team B.

3.4. Development of The Least Squares Regression Model

The dependent variable for the least squares regression model is the point spread difference between Team A and Team B. A positive point spread indicates a win for Team A and a negative point spread indicates a loss for Team A. The intercept was set to 0 when developing the model because the model should give the same absolute value score margin difference regardless of the ordering of Team A and Team B in the model. If Team A and Team B are

reversed in the model, the score margin difference will be negative in one case and in the other case positive. Differences of the four in-game statistics, the indicator variables for sets and one additional indicator variable for year were considered for possible inclusion into the model with the indicator variable for year being

$$Indicator4 = \begin{cases} 1, & if\ year\ is\ 2013 \\ 0, & otherwise \end{cases}$$

It is noted that if the game is played in 2014, this is indicated by X_8 set equal to 0.

3.5. Development of the Logistic Regression Model

We also want to develop a logistic regression model to estimate the probability that Team A wins the game based on in-game statistics, the set number in a match, and the year. The logistic regression model is also fit to the data with responses recorded as '1' for win and '0' for loss for the team of interest (Team A). The logistic regression model will estimate the probability of Team A winning the game (Abraham & Ledolter, 2006). The intercept is set to zero during the development of the logistic model for the same reason as above in the least squares model (Abraham & Ledolter, 2006). If variables, such as indicator variables for sets and year, are not significant at 0.15, they will be removed from the model.

3.6. Validation Model Development

After developing the models, we validated both models using new data. We gathered the data on volleyball matches from three universities: University of Minnesota, University of Florida, and University of Ohio, in 2015. From each of these universities, we collected data from 3 home and 3 away matches for a total 18 matches with 60 sets. First, using the data collected from each game, we put the values of the differences of the in -game statistics into the model and

estimated the point spread, $\hat{y}$, from the model to determine which team would win the game according to the model.

If $\hat{y} > 0$, Team A was predicted to win,

If $\hat{y} < 0$, Team B was predicted to win.

The results obtained from the point spread model were compared to the actual results to validate the model. If at least 70 % of the model predictions matched the actual results, we considered the model to be validated.

We also validated the logistic regression model. The data set collected from the games played in 2015 will also be used to validate this model. In this case, $\hat{y}$ is the estimated probability that Team A wins the game. Team A is predicted to win if $\hat{y} > 0.5$. If $\hat{y} < 0.5$, Team A is predicted to lose. The results obtained from the logistic model will be compared to the actual results to validate the model. If at least 70% of the model predictions are matched with the actual results, we will consider the model to be validated.

3.7. Development of the Prediction Model

After validating both models, we will attempt to use the score margin model to predict future games in which the in-game statistics are not known ahead of time. In this case, we considered a sample of matches from the universities who played matches in 2015. We randomly considered 50 matches that involve more than 20 universities. Prior to each match being played, we collected in-game statistics from all the games played by both teams from their previous two matches. The average of each of the in-game statistics was found for each team based on all games played in each of the two previous matches. Differences of the averages for each of the in-game statistics were found between the two teams and placed in the model.

We will give an example of the data collection for the prediction model. Team A played Team B. We collected data for two matches for Team A and Team B, played prior to this game. We averaged each of the in-game statistics for Team A (Table 3.4) and each of the in-game statistics for Team B (Table 3.5). Afterward, we calculated the differences between Team A and Team B for the averages of each of the in-game statistics (Table 3.6).

Table 3.4

Statistics for Team A

Team A	**Kills**	**Errors**	**PCT**	**Side-out**
Set1	15	4	30.6	56
Set2	18	4	32.6	57
Set3	16	2	33.3	61
Set1	17	4	31.1	56
Set2	11	5	13	50
Set3	10	8	6.1	43
Set4	9	5	12.1	50
Averages for Team A	13.71429	4.57	22.68	53.28

Table 3.5

Statistics for Team B

Team B	**Kills**	**Errors**	**PCT**	**Side-out**
Set1	14	4	23.3	60
Set2	19	6	23.2	63
Set3	13	7	11.5	60
Set4	16	4	27.3	68
Set1	17	4	31	76
Set2	5	12	-20.6	28
Set3	9	6	8.8	77
Set4	18	3	38.5	69
Averages for Team B	13.87	5.75	17.87	62.62

Table 3.6.

Difference between Team A and Team B on Averages of In-Game Statistics

Difference of Averages	**Kills**	**Errors**	**PCT**	**Side-out**
	-0.16071	-1.17857	4.810714	-9.33929

In Table 3.6, the value -.16071 for kills is the differences between the average number of kills for Team A and the average number of kills for Team B. The Errors, the PCT, and Side-out are similar.

CHAPTER FOUR. RESULTS

4.1. Model Development

4.1.1. Regression Model Results

First, we fit the model given in equation (4.1) based on the data collected in 2013 and 2014. Recall the dependent variable, y, was the score of Team A minus the score of Team B. All differences are in the model in the order Team A minus Team B. Recall we created indicator variables for sets 2, 3, and sets 4/5. If the indicator variables for the sets are all 0, this indicates set 1. We also created an indicator variable to indicate the year (either 2013 or 2014). If the indicator variable for the year was 0, the game was played in 2014. We will first test if the indicator variables for sets and year are significant in determining the score margin.

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\varepsilon. \quad (4.1)$$

Parameter estimates and associated t-values are given in Table 4.2. It is noted that the constant term is not significantly different than 0 and this term will be set to 0 as stated in Chapter 3. The indicator variables for the sets and the year were all non-significant at α equal to 0.15. The indicator variables will be taken out of the model.

Table 4.2 also gives the variance inflation factor for each of the estimated parameters associated with the independent variables in the model. Variance inflation factors (VIFs) can indicate multicollinearity (Abraham & Ledolter 2006). Multicollinearity exists whenever two or more of the predictors in a regression model are moderately or highly correlated. It can lead to unreliable and unstable estimates of the regression coefficients. If the value of the VIF associated with a parameter estimate is larger than 10, then we have solid evidence of multicollinearity (Abraham & Ledolter, 2006). In Table 4.2, all of the VIFs are below 10, which indicates multicollinearity is not a problem. We should be able to interpret the estimated coefficients.

We see the R-squared is equal to 91.76% (Table 4.1). This indicates that approximately 91.76% of the variation in point spread can be explained by the model.

Table 4.1

Coefficient of Determination

Root MSE	**2.06782**	**R-Square**	**0.9185**
Dependent Mean	0.68798	**Adj R-Sq**	0.9176
Coeff Var	300.56537		

The model was refit with all of the indicator variables removed and the constant term set to 0. All the variables left in the model are significant at α equal to .005. The new estimated model is given in equation (4.2). The test statistics and associated p-values are given in Table 4.3.

Table 4.2

Point Spread Model Parameter Estimates

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	**1**	0.01	0.18	0.03	0.98	0.00
Kills	**1**	0.21	0.03	7.80	0.00	2.61
Errors	**1**	-0.12	0.02	-5.17	0.00	1.76
PCT	**1**	0.02	0.01	2.81	0.01	4.22
Side-out	**1**	0.25	0.01	32.98	0.00	4.19
Set2(I1)	**1**	0.02	0.22	0.09	0.93	1.49
Set3(I2)	**1**	0.15	0.22	0.66	0.51	1.47
Set(I3)	**1**	0.14	0.24	0.58	0.57	1.40
Year2013(I4)	**1**	-0.12	0.16	-0.75	0.45	1.01

$$\hat{y}=0.21206(\text{Diff. in Kills})-0.12009(\text{Diff. in Errors})+0.01981(\text{Diff. in PCT})+0.25368(\text{Diff. in Side-out}) \quad (4.2)$$

Table 4.3

Point Spread Model Parameter Estimates

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Kills	**1**	0.21206	0.02662	7.97	<.0001
Errors	**1**	-0.12009	0.02284	-5.26	<.0001
PCT	**1**	0.01981	0.00702	2.82	0.0049
Side-out	**1**	0.25368	0.00761	33.34	<.0001

4.2. Model Development

4.2.1. Development of Logistic Regression Model

Using the same data and independent variables, we developed a logistic regression model to estimate the probability of Team A winning the game when all independent variables are given in terms of Team A minus Team B. Recall, we set the intercept equal to 0 because we want the results to be symmetric. In the logistic regression model, we tested the indicator variables for significance. The indicator variables were not significant and therefore they were removed from the model. The estimated model with significant variables is given by

$$Logit(y) = -0.1326(\text{Diff. in Kills}) + 0.0932(\text{Diff. in Errors}) - 0.0304(\text{Diff. in PCT}) - 0.2785(\text{Diff. in Side-out}) \quad (4.3)$$

Parameter estimates and associated p-values are given in Table 4.4.

Table 4.4

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Kills	1	-0.1326	0.0611	4.7038	0.0301
Errors	1	0.0932	0.0507	3.3838	0.0658
PCT	1	-0.0304	0.0172	3.1094	0.0778
Side-out	1	-0.2785	0.0303	84.5045	<.0001

4.3. Validation Model

4.3.1. Validating the Least Squares Regression Model.

In order to validate our score margin model, data was collected in 2015 from matches associated with 3 universities as mentioned in Chapter 3. Data from a total of 6 matches from each university was collected with 3 matches played at home and 3 matches played away. A total of 18 matches were considered with 55 sets.

In -game statistics were collected for each set and the differences of these in-game statistics for the two teams were put into the model. The in-game statistics collected included kills, errors, side-out percentage, and hitting percentage for each team. The estimated value for the point spread was found based on the model. This estimated value was compared to actual value.

An example of data collected from 11 sets (or games) is given in Table 4.5. The predicted margin is compared with the actual point spread. For the data given in Table 4.5, the model gave the correct team winning the game for observations 1-8, but not for observations 9-11.

Overall, considering all 53 games, the model gave the correct team winning the game for 40 games and incorrectly for 13 games with approximately 76% accuracy

Table 4.5

Validation Summary

Obs	Team A	Team B	Point Spread	Predicted Value	Kills	Errors	PCT	Side-out
1	24	26	-2	-1.503	2	4	-9	-5
2	25	21	4	2.975	6	3	1.7	8
3	25	18	7	6.912	2	-6	22.3	21
4	21	25	-4	-3.358	-3	0	-9.4	-10
5	10	15	-5	-5.972	1	4	-19.1	-21
6	25	15	10	8.722	2	-7	30.8	27
7	25	13	12	10.240	5	-7	36.9	30
8	25	11	14	12.312	6	-7	41.2	37
9	25	16	9	-7.779	1	4	-21	-28
10	26	24	2	-1.027	3	2	-7.8	-5
11	21	25	-4	2.367	-2	-1	6.8	10

4.3.2. Validating the Logistic Regression Model.

We used the same data that was collected to validate the least squares regression model (or score margin model) in order to validate the logistic regression model. In-game statistics were collected for each set and the differences of these in-game statistics between the two teams were put into the model. The estimated probability of Team A winning the game was found. If the estimated probability was greater than 0.50, Team A was predicted to win; otherwise, Team A was predicted to lose. This was compared to the actual results. We predicted the correct team winning the game in 39 out of 53 games for the logistic model. The model gave the correct result in 74% of the cases. Since this percentage is greater than 70%, we considered the model validated.

4.4 Prediction Model

The least squares model or score margin model was used for prediction. As stated in Chapter 3, a random sample of 50 matches from more than 20 universities was collected. For the two universities in a match, the in-game statistics based on all the games in the two previous matches were averaged for each university. The differences of these averages for each of in-game statistics were placed in the score margin model. If the score margin model gave a positive result, this indicates the model predicted Team A to win a game when playing Team B and therefore, overall, Team A should win the match.

If the model predicted the score margin of 3, on the average we would expect Team A to win a game by 3 points and therefore we would predict Team A to win the match. If the model predicted a score margin of negative 2, on the average we would expect Team B to win a game by 2 points and therefore, we would predict Team B to win the match. An example taken from 6 matches is given in Table 4.6.

Table 4.6

Prediction and Actual Results

# of Games won by Team A	# of Games won by Team B	Predicted Score Margin	Team Predicted to Win Match	Kills	Errors	PCT	Sideout
1	3	-2.17	B	-0.03	0.14	0.10	-2.37
2	3	0.55	B	-0.18	-0.11	-0.02	0.86
2	3	-2.62	B	-0.14	-0.22	-0.19	-2.07
1	3	-0.29	B	0.42	-0.08	0.11	-0.74
3	0	3.23	A	0.16	0.18	0.10	2.80

The overall results are given in Table 4.7.

Table 4.7

Actual vs Predicted Results

	Actual	Predicted Results
Win	27	34
Loss	23	16
Total	50	50
Overall Accuracy		68%

We correctly predicted 68% of the matches. This is comparable to results for football in Long and Magel (2013) and hockey in Roith and Magel (2014).

CHAPTER FIVE. CONCLUSION

Two models were developed for use with NCAA Division I Women's Volleyball games. We developed one point spread model for a game or set that explained the variation in point spread of a women's volleyball game or set based on knowing the differences between the two teams' in-game statistics. The second model was a logistic regression model that estimated the probability of Team A winning the game if the differences of in-game statistics were known. Both models were validated in CHAPTER FOUR. If the actual in-game statistics were known, the least squares regression model had an estimated accuracy of 76% to correctly predict the results. If the actual in-game statistics were known, the logistic regression model had an estimated accuracy of 74%.

The score margin model was used to predict the results for 50 matches played in 2015 Division I Women's Volleyball. Average in-game statistics from the past two matches for each team were found. The differences of these averages were placed in the model in place of the actual in-game statistics. If the model gave a positive result, Team A was predicted to win. Otherwise, Team B was predicted to win. The score margin model had an accuracy of 68% when these averages were used. Overall, when the averages from past matches were used, we were able to correctly predict 68% of the games, which was better than just flipping a coin. In order to improve this accuracy in the future, perhaps additional in-game statistics could be found which help to further explain the point margin in a volleyball game.

REFERENCES

Abraham, Bovas, and Ledolter, Johannes (2006). Introduction to Regression Modeling. Belmont:Thomson/Cole, 2006 Print.

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

Long, Joe and Magel, Rhonda (2013). "Identifying Significant In-Game Statistics and Developing Prediction for Outcomes of NCAA Division I Football Championship Subdivision (FCS) Games". Journal of Statistical Science and Application, Vol.1, N. 1, pp 51-62 (December 2013).

Melynkov, Yana and Magel, Rhonda (2014). "Examining Influential Factors and Prediction Outcomes in European Soccer Games". International Journal of Sport Science, Vol 4, No. 3, 2014.

NCAA.2013. DI Women's Volleyball Championship. Retrieved from http://www.ncaa.com/2013-division-I-womens-volleyball-tournament-stats on 10/03/2016.

NCAA.2014a. The National Collegiate Athletic Association "2014 and 2015 NCAA Women's Volleyball Rules and Interpretations", 2014.

NCAA. 2014b. DI Women's Volleyball Championship. Retrieved from http://www.ncaa.org/championships/statistics/2016-ncaa-womens-volleyball-records-book on 10/03/2016.

NCAA. 2014c . DI Women's Volleyball Championship. Retrieved from http://www.ncaa.com/2014-division-I-womens-volleball-tournamnet-stats on 10/03/2016.

NCAA. 2015. DI Women's Volleyball Championship. Retrieved from http://www.ncaa.com/2014-division-I-womens-volleball-tournamnet-stats on 10/03/2016.

Roith, Joe and Magel, Rhonda (2014). "An Analysis of Factors Contributing to Wins in the National Hockey League" International Journal of Sport Science, Vol 4, No. 3, 2014.

Unruth, Sam and Magel, Rhonda (2013). "Determining Factors Influencing the Outcome of College Basketball Games". Open Journal of Statistics. Vol, 3 No.4 (August 2013).

Wang, Wenting, and Magel, Rhonda (2014). “Predicting Winner of NCAA Women’s Basketball Tournament Games”. International Journal of Sport Science, 2014, 4(5):173-180.

APPENDIX A. VALIDATION DATA

Team A	Team B	Point Spread	Predicted value	Kills	Errors	PCT	Side-out
24	26	-2	-1.5	2	4	-9	-5
25	21	4	2.98	6	3	1.7	8
25	18	7	6.92	2	-6	22.3	21
21	25	-4	-3.36	-3	0	-9.4	-10
10	15	-5	-5.98	1	4	-19.1	-21
25	15	10	8.73	2	-7	30.8	27
25	13	12	10.24	5	-7	36.9	30
25	11	14	12.32	6	-7	41.2	37
23	25	-2	-0.37	3	-1	7.2	-5
25	14	11	10.12	3	-8	33.3	31
25	20	5	4.25	6	2	8.6	12
26	28	-2	-1.36	-3	-1	-4.1	-3
15	13	2	2.55	2	-1	11.4	7
25	23	2	0.23	-5	-3	-4.1	4
25	20	5	3.27	3	2	4.3	11
25	21	4	3	-3	-5	12.2	11
25	23	2	1.34	0	-2	4.4	4
19	25	-6	-4.43	-2	2	-10.6	-14
26	24	2	1.16	-1	-1	-1	5
25	22	3	2.04	-1	-3	5.7	7
25	16	9	7.45	4	0	12.9	25
25	23	2	0.79	-3	0	-4.8	6
25	21	4	2.8	2	0	4.7	9
25	15	10	11.91	12	-3	57.3	31
25	9	16	14.97	5	-10	13.8	49
25	15	10	9.1	4	-6	34.5	27
22	25	-3	0.74	-6	-1	18.4	6
25	16	9	-7.78	1	4	-20.7	-28
26	24	2	-1.03	3	2	-7.8	-5
21	25	-4	2.37	-2	-1	6.8	10
16	18	-2	2.5	2	-2	3	7
25	22	3	2.39	-1	-5	11.6	7
25	20	5	4.97	5	-2	18.7	13
25	13	12	11.46	6	-6	42.2	34
25	21	4	4.71	12	5	11.5	10

Team A	Team B	Point Spread	Predicted value	Kills	Errors	PCT	Side-out
15	25	-10	-9.32	-10	1	-36.9	-25
17	25	-8	-7.23	2	7	-23.5	-25
15	25	-10	-10.8	-8	6	-39	-30
29	31	-2	-1.72	-4	-2	-4.9	-4
16	25	-9	-7.66	-6	4	-29.2	-21
15	25	-10	5.93	-6	3	35.7	27
16	25	-9	6.94	-3	4	22.3	30
25	19	6	-4.14	1	-2	-13.9	-17
14	25	-11	7.99	-5	5	38.9	35
25	11	14	13.9	5	-6	35.6	45
25	20	5	4.08	4	0	9.2	12
25	21	4	2.54	6	3	5.5	6
25	12	13	-10.29	11	1	-29	-47
23	25	-2	1.46	3	0	-9.8	4
25	23	2	-1.04	0	1	4.8	-4
25	18	7	-2.71	5	-2	-10.4	-15

APPENDIX B. PREDICTION DATA

# of Games won by Team A	# of Games won by Team B	Predicted Score Margin	Team Predicted to Win match	Kills	Errors	PCT	Side-out
1	3	-2.17	B	-0.03	0.14	0.1	-2.37
2	3	0.55	B	-0.18	-0.11	-0.02	0.86
2	3	-2.62	B	-0.14	-0.22	-0.19	-2.07
1	3	-0.29	B	0.42	-0.08	0.11	-0.74
3	0	3.23	A	0.16	0.18	0.1	2.8
1	3	0.04	B	0.09	-0.02	-0.04	0
3	2	-1.58	A	-0.43	0.04	-0.04	-1.16
3	0	4.18	A	0.06	0.23	0.23	3.65
3	1	0.49	A	-0.1	0.1	0.08	0.41
3	1	4.96	A	0.48	0.29	0.48	3.7
3	1	3.16	A	0.4	0.11	0.14	2.52
3	1	3.86	A	0.27	0.05	0.03	3.52
1	3	-1.83	B	-1.11	0.03	-0.09	-0.66
0	3	-0.33	B	-0.61	0.02	-0.14	0.4
0	3	-4.31	B	-0.42	-0.3	-0.28	-3.32
3	0	-2.79	A	0.36	-0.03	-0.07	-3.05
3	1	1.66	A	0.57	0.39	0.34	0.35
3	1	2.76	A	0.29	-0.18	0.15	2.49
0	3	1.22	B	-0.1	-0.04	0.01	1.35
2	3	2.43	B	0.31	0	0.12	2.01
0	3	-6.72	B	-1.33	-0.15	-0.4	-4.84
0	3	-0.41	B	-0.74	0.05	-0.11	0.39
3	0	3.07	A	1.04	-0.23	0.14	2.11
3	0	7.47	A	0.64	0.66	0.68	5.5
0	3	-6.98	B	-0.99	-0.34	-0.48	-5.17
3	0	-1.6	A	-0.4	-0.33	-0.24	-0.63
0	3	2.26	B	0.94	0.14	0.2	0.98
0	3	0.74	B	0.11	0.4	0.27	-0.03
0	3	-4.19	B	-0.75	0.1	-0.04	-3.51
3	2	3.61	A	0.41	-0.14	0	3.33
3	0	-0.37	A	0.12	-0.06	0.02	-0.45
3	1	1.1	A	0.18	0	0.02	0.9

# of Games won by Team A	# of Games won by Team B	Predicted Score Margin	Team Predicted to Win match	Kills	Errors	PCT	Side-out
3	2	-0.56	A	-0.21	-0.16	-0.15	-0.04
3	2	-3.21	A	-0.32	-0.19	-0.22	-2.48
1	3	1.52	B	0.14	0.2	0.12	1.06
3	0	7.28	A	0.57	0.4	0.59	5.72
3	0	7.54	A	0.54	0.38	-0.09	6.71
3	1	-1.11	A	-0.66	0.11	-0.18	-0.39
2	3	-8.16	B	-0.79	-0.06	-0.36	-6.95
1	3	5.82	B	0.52	-0.13	0.08	5.35
3	0	4.39	A	0.59	0.34	0.32	3.14
1	3	-2.72	B	-0.18	-0.12	-0.25	-2.17
3	1	-2.11	A	-0.22	-0.31	-0.24	-1.34
3	0	3.18	A	0.94	0.1	0.24	1.9
3	0	10.64	A	1.06	0.35	0.62	8.61
3	1	4.4	A	0.87	0.08	0.28	3.17