IJSRM
INTERNATIONAL JOURNAL OF SCIENCE AND RESEARCH METHODOLOGY
An Official Publication of Human Journals

Human Journals
Research Article
February 2018 Vol.:8, Issue:4

Explaining and Estimating the Point Spread of a Women's NCAA Volleyball Game

IJSRM
INTERNATIONAL JOURNAL OF SCIENCE AND RESEARCH METHODOLOGY
An Official Publication of Human Journals

Rhonda Magel*, Deling Zhang, Ronald Degges

Department of Statistics North Dakota State University
Fargo, ND 58108

Submission: 24 January 2018
Accepted: 29 January 2018
Published: 28 February 2018

HUMAN JOURNALS

www.ijsrm.humanjournals.com

Keywords: Ordinary Least Squares Regression; Volleyball, Kills, Side-out, Errors, Hitting Percentage, Matches

ABSTRACT

A model is developed to help explain the point spread of a Women's Division 1 volleyball game using matches played over a two-year period of time-based on four in-game statistics: difference in kills; the difference inside-out percentages; the difference in hitting percentages; and the difference in the number of errors. The model is validated on a different data set, and then the model is used for prediction replacing in-game statistics with past averages for both volleyball teams playing in each match in the data set.

INTRODUCTION

The game of volleyball was introduced by William G. Morgan in 1895 and was originally called "mintonette", [1]. Today, volleyball has become a very popular sport and it ranks second only to soccer as to the number of people who participate in it. A 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 try to do this so that the other team can't return it but touches the ground of the opposing side [1].

In this research, we are interested in College NCAA Division 1 games of Women's Volleyball. We would like to develop a model that explains the point spread of a women's volleyball game. Once developed, the point spread model will give coaches and players an idea as to how much weight is given to certain factors in a volleyball game so that the teams will have an idea as to what skills to spend more of their time working on. The second objective of this research is to use this model to try and predict future outcomes of NCAA Women's Division 1 volleyball matches.

There have been many papers published modeling the outcome of sporting games, including basketball, football, hockey, and soccer [2], [3], [4], [5], [6], [7], [8], [9] among others. The vast majority of publications have been geared to modeling or predicting the outcomes in men's sporting games. A few publications, such as Wang and Magel [8] have developed models for explaining the outcome of women's sporting game, in this case, basketball. There have been few publications, if any, for modeling the outcome of a women's volleyball game. Researchers have discussed the importance of each skill in regular and beach volleyball games including both men's and women's games ([10], [11], [12]). Marcelino et al. [13] investigated the home court advantage of a men's volleyball match in relation to the set number.

The scoring for women's volleyball has changed over time. Originally, the serving team could only earn points. A point was earned when the team not serving the ball could not return the ball and it hit the ground. Games went until one team had at least 15 points and had at least 2 points more than the other team. If a team reached 15 points but was less than 2 points ahead of the other team, the game kept going until a distance of 2 points was obtained. Scoring has now changed so that a point is scored after every time the ball is served. Essentially, teams score points whenever the other team makes a mistake, and a point is awarded on every serve, called volley scoring. The game now goes until a team has scored at least 25 points and has a score of at least 2 more points than the opposing team. In NCAA Women's Volleyball, teams play a match. A team wins a match when they are the first to win 3 games. Matches consist of 3 to 5 games or sets. [14].

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 match ahead of time in future matches. A volleyball match consists of 3 to 5 games or sets. The first team to win 3 sets wins the match.

MATERIALS AND METHODS

We considered the following four in-game statistics when developing the model because several teams keep these: number of kills, number of attack errors, hitting percentage, and side-out percentage. Other in-game statistics do exist but were not available for the majority of women's volleyball games

A "kill" is awarded to a team if the ball is hit into the other team's court and the result is that the ball is not returnable. A team is charged with an error if they attempt to hit the ball in the other team's court so that it is not returnable, but the ball lands in their court instead. Hitting percentage is calculated for a team as being the number of kills minus the number of errors divided by the number of attack attempts, where an attack attempt is an attempt to hit the ball in the other team's court so that it is not returnable. The side-out percentage is calculated by dividing the serve receive points by the number of serves receive attempts, times 100. [[1].

Data were collected from a sample of 108 matches involving 18 universities for two seasons (2013 and 2014) of NCAA Women's Division I volleyball games, where each match consisted of 3 or more games or sets, in the years of 2013 and 2014. Data were collected from 3 home matches and 3 away matches for each of the 18 universities. The number of kills in the data collected varied from 12 to 19, the number of errors varied from 2 to 8, the hitting percentage varied from 16.7 to 40, and the side out percentage varied from 51 to 77. Differences for each of these variables were found between the two teams playing a game in the order Team A minus Team B

Tables 1 and 2 are examples of the collected data that show the values for the in-game statistic for Teams A and B from each set in the match. We want to estimate the score margin or point spread of a game; namely points scored by Team A minus points scored by Team B.

Table 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 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 dependent variable in the model is point spread. 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 [14]. The differences are in the order Team A-Team B. The point spread and differences between each of the in-game statistics are given in Table 3.

Table 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

In addition to the in-game statistics, three indicator variables are considered included in the model. These new variables indicate the number of the set or game played in the match. A match may consist of 3, 4, or 5 sets. The indicator variables for the sets were defined as:

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

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

The constant term in the model should be zero since the order of the teams in the model should not matter.

After developing the model, we validated the model using new data. We gathered the data on volleyball matches from three universities: the University of Minnesota, University of Florida, and the University of Ohio in 2015 [14]. 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 indicated to win,

If $\hat{y} < 0$, Team B was indicated 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 indications matched the actual results, we would consider the model to be validated.

After validating the model, we attempted to use the model to predict future games, and ultimately future matches, in which the in-game statistics are not known ahead of time. In this case, we considered a sample of matches 50 from more than 20 universities who played matches in 2015. The average of each of the in-game statistics was found for both teams playing in a match based on all games a team played in each of their two previous matches. Differences in the averages for each of the in-game statistics were found between the two teams involved in the match and placed in the model. If the model estimated a positive point spread, that would indicate that the model was predicting Team A to win the game, and ultimately the match. If the model estimated a negative point spread, the model was indicating that Team B would win the game and ultimately the match.

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 4) and each of the in-game statistics for Team B (Table 5).

Afterward, we calculated the differences between Team A and Team B for the averages of each of the in-game statistics (Table 6). In Table 6, the value -.16071 for kills is the difference 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 percentages are similar.

Table 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	4.571429	22.68571	53.28571

Table 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	13.875	5.75	17.875	62.625

Table 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

RESULTS

Model Development

First, we fit a regression model based on the data collected in 2013 and 2014. Recall the dependent variable, y, was the score of Team A minus 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 first tested if the indicator variables for sets and year are significant in determining the point spread.

Parameter estimates and associated t-values are given in Table 7. It is noted that the constant term is not significantly different than 0 and this term will be set to 0 as earlier stated. 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 7 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 be used to indicate multicollinearity [15]. If the value of the VIF associated with a parameter estimate is larger than 10, then we have solid evidence of multicollinearity [15]. In Table 7, all of the VIFs are below 10, which indicates multicollinearity is not a problem. We should be able to interpret the estimated coefficients.

The adjusted R-squared for the model is equal to 91.76%. This indicates that approximately 91.76% of the variation is the model can explain point spread.

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 estimated model is given by equation (Eq1).

Table 7: 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
Kill	**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

$$\hat{y}=0.21206(\text{Diff. in Kills})-0.12009(\text{Diff. in Errors})+0.01981(\text{Diff. in PCT})$$
$$+0.25368(\text{Diff. in Side-out}) \qquad \text{(Eq 1)}$$

Validating the Model

In order to validate our score model, data were collected in 2015 from matches associated with 3 universities as mentioned earlier. Data from a total of 6 matches from each university was collected with 3 matches played at home and 3 matches played away. 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 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 8. The predicted margin is compared with the actual point spread. For the data given in Table 8, 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 8: Validation Summary

Obs	Team A	Team B	Point Spread	Predicted Value	Kill	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

Using the Model for Prediction 3

The least squares model was used for prediction. 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 both of the universities. The differences in these averages for each of in-game statistics were placed in the point spread model. If the point spread 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 a point spread 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 point spread 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 9.

The model correctly predicted 34 out of the 50 matches. This gave an accuracy of 68%.

Table 9: Prediction and Actual Results

# of Games won by Team A	# of Games won by Team B	Predict Point Spread	Team Predict to Win Match	Kills	Errors	PCT	Sideout
1	3	-2.17	B	-0.03	0.14	0.10	-2.37
2	3	-1.58	B	-2.02	-0.36	-1.95	-4.57
3	1	3.16	A	1.88	-0.92	7.12	9.92
0	3	0.74	A	0.54	-3.30	13.39	-0.13
3	2	3.61	A	1.93	1.14	-0.03	13.14

CONCLUSION

A model was developed that explained 92% of the variation in point spread of a volleyball game when four in-game statistics were known. From the model, we can determine how much weight is put on each of the four variable differences in determining the point spread of a volleyball game. For example, if a team increases its side-out percentage to 4 percentage points above the team they are playing, this gives the team approximately a one-point advantage. A difference in one kill amounts to about a 0.2-point advantage. The model was also used to predict outcomes of future matches. To do this, average in-game statistics of both team playing in a match were found based on the former two matches that each team played. The differences of these averages were placed in the model, and the model was used to predict which team would win the game and ultimately the match. We were able to correctly predict the winning team in 34 out of the 50 matches, for an accuracy of 68%. This is comparable to results for football in Long and Magel [5] and hockey in Roith and Magel [16].

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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).
[7] Melynkov, Yana and Magel, Rhonda (2014). "Examining Influential Factors and Prediction Outcomes in European Soccer Games". International Journal of Sports Science, Vol 4, No.3, 2014.
[8] Wang, Wenting, and Magel, Rhonda (2014). "Predicting Winner of NCAA Women's Basketball Tournament Games". International Journal of Sports Science, 2014, 4(5):173-180.
[9] Shen, Gang, Gao, Di, Wen, Qian, & Magel, Rhonda (2016). "Predicting Results of March Madness Using Three Different Methods", Journal of Sports Research, Vol. 3, Issue 1, pp 10-17.
[10] Palao, J.M., Santos, J.A. & Ureña, A. (2004). Effect of team level on skill Performance in Volleyball, International Journal of Performance Analysis in Sport, Vol.4, Issue 2.
[11] Giatsis, George (2008). Statistical Analysis of Men's FIVB Beach Volleyball Team Performance. International Journal Of Performance Analysis in Sport, 31-43.
[12] Miskin, M., Fellingham, G. & Florence, L. (2010). Skill Importance in Women's Volleyball. *Journal of Quantitative Analysis in Sports*, 6(2), Retrieved 5 Dec. 2017, from doi:10.2202/1559-0410.1234
[13] Marcelino, R., Mesquita, I., Andres, J., and Sampaio, J. (2009). "Home advantage in high-level volleyball varies according to set number", Journal of Sports Science & Medicine, 8(3), pp. 352-356.
[14] NCAA. 2014b. DI Women's Volleyball Championship. Retrieved from http://www.ncaa.org/championships/statistics/womens-volleyball-statistics on 10/03/2016.
[15] Abraham, Bovas, and Ledolter, Johannes (2006). Introduction to Regression Modeling. Belmont: Thomson/Cole, 2006 Print.
[16] Roith, Joe and Magel, Rhonda (2014). "An Analysis of Factors Contributing to Wins in the National Hockey League" International Journal of Sports Science, Vol 4, No. 3, 2014.