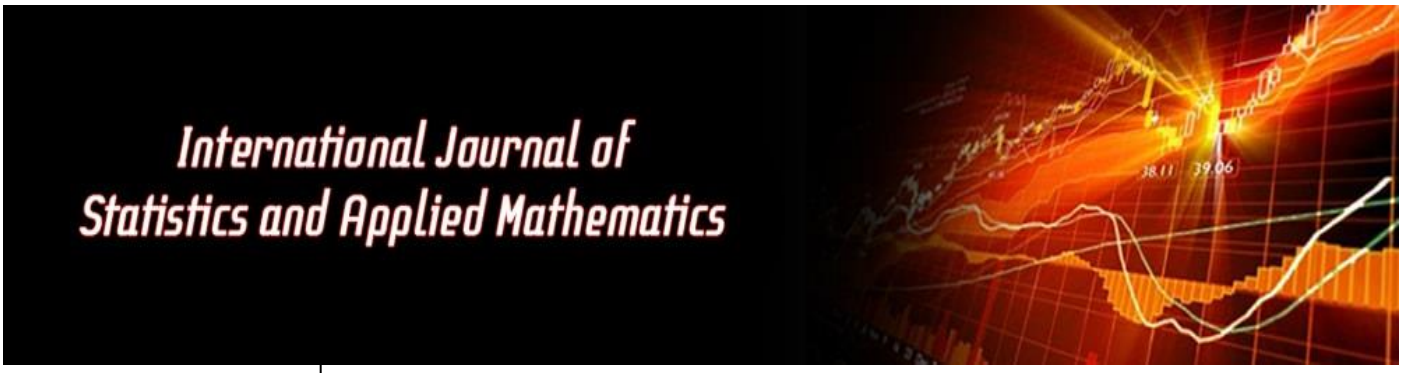
International Journal of Statistics and Applied Mathematics 2019; 4(2): 16-21



ISSN: 2456-1452
Maths 2019; 4(2): 16-21

www.mathsjournal.com
Received: 04-01-2019
Accepted: 07-02-2019

Wenting Wang
Department of Statistics,
North Dakota State University,
Fargo, ND 58108, USA

Rhonda Magel
Department of Statistics,
North Dakota State University,
Fargo, ND 58108, USA

Correspondence
Wenting Wang
Department of Statistics,
North Dakota State University,
Fargo, ND 58108, USA

Bracketing NCAA women's volleyball tournament

Wenting Wang and Rhonda Magel

Abstract
Models were developed using seasonal averages, and ranks of seasonal averages, to predict winners of the various rounds of the NCAA Division 1 Women's Volleyball tournament. Data used to develop these models was collected during 2011-2013. These models were used to predict winners of the 2015 tournament matches and the overall winner. A match statistic model was also developed using data from actual matches played in the NCAA Women's Tournament in 2015. The match statistic model was used to predict winners of the matches in the 2016 tournament and the overall winner. It was found that models based on using seasonal averages performed better at predicting the winners of the matches and then the overall tournament than the models developed using match statistics replacing match statistics with seasonal averages of the corresponding statistics. Results are given.

Keywords: Least squares regression, validating, predicting, stepwise regression

1. Introduction
The NCAA Division 1 Women's Volleyball Tournament has been held annually since 1981. During the first year, only 20 universities competed in the tournament. The tournament gradually expanded until it reached its current size of 64 universities [1]. There are 330 universities that have NCAA Division 1 Women's Volleyball Teams, and 64 teams end up competing in the tournament. Of the 64 teams competing in the tournament, 32 teams will receive automatic bids and a committee [2] selects the remaining teams. The tournament is single-elimination. After the first round, there will be 32 teams remaining for the second round, 16 teams remaining for the third round, 8 teams remaining for the 4th round, 4 teams in the 5th round, and then 2 teams for the 6th and final round [2].
The purpose of this research is to develop models that help predict the winner of each match in each of the rounds of the tournament using seasonal averages of various volleyball statistics, and then using ranks of seasonal averages of various volleyball statistics. We will validate the models using a different data set and then use the models to predict winners of each match in the 2015 tournament as well as the overall winner of that tournament. We will next develop a model to explain the differences in number of games won by teams in a volleyball match based on actual match statistics. Data for this model comes from matches played in the 2015 tournament. This model will be validated and then used to predict winners in the various rounds in the 2016 tournament replacing the significant match statistics with the corresponding seasonal averages for the teams playing in the tournament.
There has not been much research into predicting winners of volleyball games or matches. We will mention two previous studies. Giatsis (2008) [3] conducted an analysis on men's beach volleyball. The purpose of his study was to determine the differences in playing characteristics between winning and losing teams in FIVB Men's Beach Volleyball World Tour Tournament. Giatsis used independent t-tests and a discriminant function analysis to determine which skills contributed significantly to winning in matches. He found the opponents' attack errors was the most significant factor contributing to winner's win.
Zhang (2016) [4] developed a multiple linear regression model using in-game statistics that explain the point spread of a volleyball game and a logistic regression model that estimates the probability of a team winning the game based on the in-game statistics for college women's volleyball games. The point spread model was used to predict the results of future volleyball games by replacing the in-game statistics with the averages of the in-game statistics based on

the past two previous matches of both teams playing each other. This research did not involve predicting winners of a tournament.

2. Methods used for Seasonal Averages Models

We collected data from three years of the NCAA Division 1 Women's Volleyball Tournament, 2011-2013, to develop our first set of models. Ranks of seasonal averages and seasonal averages were found for all the teams in each of these tournaments on the variables given in Table 1 from the website NCAA.COM [5].

Table 1: Variables in consideration for seasonal average and rank.

Variables in consideration	Definitions
Aces Per Set	A serve that results directly in a point when a player attempts to serve the ball over the net into the opponent's court for each set. [6]
Assists Per Set	When a player passer, sets or digs ball to teammate who gets a kill for each set. [6]
Blocks Per Set	Player(s) block leads directly to a point for each set. [6]
Digs Per Set	When a player receives an attacked ball and keeps the ball in play for each set. [6]
Hitting Percentage	Hitting Percentage = (Total kills – Total Errors)/ Total Attempts. [6]
Kills Per Set	An attack that directly leads to a point for each set. [6]
	Match W-L Percentage = Numbers of games won / Total sets played. [6] Note: The value for Match W-L Percentage will between 0 to 1.

The seasonal averages and ranks of teams for each of the variables and for each year were based on all games in the season, up to, but not including any game in the tournament. There were 192 teams playing 96 matches in the first rounds of the tournaments during the years 2011-2013. Two models to predict the outcome of a first round volleyball match were developed. The independent variables in the model were the differences of ranks of the seasonal averages of the variables given in Table 1 between the two teams playing in a match in the order the "team of interest" minus "opposing team". When collecting the data, the "team of interest" in half of the first round games was the stronger team (higher seeded), and in the other half, the "team of interest" was the weaker team (lower seeded). The dependent variable was the difference in the number of games won between the two teams in a match.

The independent variables considered for the second model were the differences of the seasonal averages between the two teams playing in the match of the variables given in Table 1. Least squares regression using the stepwise selection procedure was used for both models with α equal to 0.1 for entry and exit into the models [7]. The intercept term was set to zero in both models, because if the seasonal averages or ranks of the seasonal averages between the two teams were the same, the differences in number of games won should be zero.

Two models were next developed based on the second round matches of the tournaments. The last set of two models were based on the third and higher round matches in the tournaments. The independent variables considered for entry into the model were the same as before.

Only the models for each round resulting from using seasonal averages of the statistics are given. Models based on ranks of seasonal averages did not perform as well. The model based on data in round 1 is given by Equation 1.1. The summary of the stepwise selection process when developing the model is given in Table 2. The model explains 39.46% of the variation in differences of games won in round 1 of the NCAA Division 1 Women's Volleyball Tournament. The model for round 2 matches is given in Equation 2.2, and the model for round 3 and higher matches is given in Equation 2.3. Tables 3 and 4 are associated with the round 2 and round 3 and higher models. The R^2 associated with each model is given.

$\hat{Y}$ = (-2.04026* Diff_ Aces) + (28.72233* Diff_Hitting%) Equation 2.1 (1st)

Table 2: Summary of Stepwise Selection for First- Round Model based on Averages

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Diff_Hitting%		1	0.3080	0.3080	17.4884	42.28	<.0001
2	Diff_Aces		2	0.0866	0.3946	5.5395	13.44	0.0004

$\hat{Y}$ = -0.57886* Diff_Digs, Equation 2.2 (2nd)

Table 3: Summary of Stepwise Selection for Second-Round Model based on Averages

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Diff_Digs		1	0.1709	0.1709	0.3116	9.69	0.0032

$\hat{Y}$ = 7.33912* Diff_Match_W-L%, Equation 2.3 (3rd)

Table 4: Summary of Stepwise Selection for Third and Higher Round Model Based on Averages

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Diff_Match_W-L%_		1	0.1547	0.1547	0.6923	8.06	0.0068

2.1 Validating first round model using seasonal averages

Data from the first round of the 2014 tournament was used to validate the first round model. Table 5 gives the results as to how well the model did at predicting the winners of Round 1 games. The model had an overall accuracy of 65.6%.

Table 5: Accuracy of least squares regression model developed by seasonal averages when validating first round of 2014

Point spread		Predicted		
		Win	Loss	Total
Actual	Win	10	4	14
	Loss	7	11	18
	Total	17	15	32
Overall Accuracy				65.6%

2.2 Validating second round model using seasonal averages

Data from the second round of the 2014 tournament was used to validate the second round model. Table 6 gives the results as to how accurately the least squares regression model for second round of the NCAA 2014 women's volleyball tournament. The model had an overall accuracy of 68.8% of correctly predicting winners of matches in the second round.

Table 6: Accuracy of least squares regression model developed by seasonal averages when validating second round of 2014

Point spread		Predicted		
		Win	Loss	Total
Actual	Win	4	3	7
	Loss	2	7	9
	Total	6	10	16
Overall Accuracy				68.8%

2.3 Validating third and higher rounds using seasonal averages

Data from the third and higher rounds of the 2014 tournament was used to validate the third and higher rounds least squares regression model. Table 7 gives the results as to how accurately the model for third and higher rounds of the NCAA 2014 women's volleyball tournament. This model had an overall accuracy of 53.3%.

Table 7: Accuracy of least squares regression model developed by seasonal averages when validating third and higher rounds of 2014

Point spread		Predicted		
		Win	Loss	Total
Actual	Win	5	3	8
	Loss	4	3	7
	Total	9	6	15
Overall Accuracy				53.3%

2.4 Bracketing the 2015 tournament before tournament begins – Prediction

Results were predicted for every round before the 2015 tournament began. Seasonal averages of significant difference variables were found for all teams playing in the first round and put into first round model and first round winners predicted. Seasonal averages of significant difference variables for each set of teams predicted to play each other in the second round were placed in second round model and winners of this round were predicted. Seasonal averages of significant difference variables of teams predicted to play each other in the third round were placed in the third round model and winning teams predicted for this round. This process continued until the overall winning team was predicted. The predicted results as to which team won were compared against the actual results for each round in the 2015 tournament.

An example will be given as to how the model for each round was used in 2015 tournament. Southern California played Cleveland State in the first round of the 2015 tournament. In the regular season, Southern California averaged 1.52 aces per game and Cleveland State averaged 1.05 aces per game for a difference of 0.47. South California had an average hitting percentage of 0.292 and Cleveland State had an average hitting percentage of 0.248 for a difference of 0.044. When these differences were placed in the first round model given in Equation 2.1, the result is given by:

$\hat{y} = (-2.04026* 0.47) + (28.72233* 0.044) = 0.3$

Since $\hat{y} > 0$ this match was coded as correctly predicting Southern California to win the match. Southern California did win the match by 3 games to 1.

Round 1
Number correct: 25
Number incorrect: 7
Total: 32

In the second round of the 2015 tournament, BYU played Western Kentucky. The average number of digs per game for the season for BYU was 14.63, and for Western Kentucky, it was 14.96, for a difference of -0.33.

Using the second round model given in Equation 2.2, the following result was obtained:

$\hat{y} = -0.57886* -0.33 = 0.19$

Since $\hat{y} > 0$ this match was coded as a correctly predicted win for BYU, who won the match by 3 games to 0.

Round 2
Number correct: 8
Number incorrect: 8
Total: 16

Texas played Florida in the fourth round of the 2015 tournament. Texas had an average win-loss percentage for the season of 0.929 and Florida had a win-lost percentage of 0.793, for a difference of 0.136. Using the model for the third and higher rounds given in Equation 2.3, the following result was obtained:

$\hat{y} = 7.33912* 0.136 = 0.99$

Since $\hat{y} > 0$ this match was correctly predicted as a win for Texas, who won the match by 3 games to 2.

Round 3-6
Number correct: 7
Number incorrect: 8
Total: 15

Only the teams predicted to play each other in the second round were used in the model. This was also true in other rounds. It was possible that the model predicted a win for a team in a round that did not even make it to that round.

2.5 Results for prediction by using models developed by difference of seasonal averages

In 2015, a continuous process was used in verifying the models instead of doing round by round predictions as in

2014. In other words, a complete bracket was filled out in 2015 before any game was played.

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

Accuracy of least squares regression model results was given in Table 8.

Table 8: Prediction results of each round for 2015: (Least squares regression models developed by seasonal averages)

	Correct	Incorrect	Total matches
First round	25	7	32
Second round	8	8	16
Third round	4	4	8
Fourth round	2	2	4
Fifth round	1	1	2
Final round	0	1	1
Overall Accuracy			63.5%

3. Method Used to Develop model by using actual match statistics

A model was also developed to explain the differences in number of games won in a volleyball match based on known match statistics. In order to develop this model, data was collected from the 2015 NCAA women's volleyball tournament matches on the variables given in Table 9 for both teams playing in a match and the differences were found. The independent variables considered for inclusion into the model were these differences. The dependent variable was the difference in number of games won in a match between the two teams. Least squares regression using the stepwise selection procedure [7] was used to determine which variables under consideration should be in the model.

Table 9: Variables in consideration for match statistics

Variables in consideration	Definitions
Attack Kill	An attack that directly leads to a point. [6]
Attack Error	An attack that directly results in a point for the opposing team. [6]
Attack Percentage	Attack Percentage = (Total kills – Total Errors)/ Total Attempts. [6]
Serve SA (Service ace)	A service ace (SA) is a serve that results directly in a point when a player attempts to serve the ball over the net into the opponent's court. [6]
Serve (Reception Error)	When a result for a point for the opposing team a player of team must be charged with a reception error. [6]
Digs	When a player receives an attacked ball and keeps the ball in play. [6]
Blocks	Player(s) block leads directly to a point. [6]

The model to help explain the variation in differences of games won for each match in first round through final round based on using differences between match statistics of the significant variables was developed and found to be:

$\hat{Y}$ = (0.04538* Diff_AttackK) + (8.12106 * Diff_Attack%) + (0.21009* Diff_ ServeSA)

Equation 3

The standard errors and p-values associated with each of the parameter estimates for the model are given in Table 10. Table 11 gives the steps associated with the stepwise selection technique. The model with the 3 significant variables explains an estimated 82% of the variation in the differences in number of games won in a match.

Table 10: Match Model Parameter Estimates

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Attack_K	1	0.04538	0.02365	1.92	0.0634	2.40263
Attack_PCT	1	8.12106	1.71816	4.73	<.0001	2.43370
SERVE_SA	1	0.21009	0.04758	4.42	<.0001	1.19471

Table 11: Summary of Stepwise Selection for Match Statistics Model

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Attack_PCT		1	0.6883	0.6883	24.1946	79.48	<.0001
2	SERVE_SA		2	0.1154	0.8036	4.2859	20.57	<.0001
3	Attack_K		3	0.0192	0.8228	2.6413	3.68	0.0634

3.1 Validating the match statistics model

Using the model developed with match statistics, the differences in the number of games won for each set of teams playing in a match for each of the 63 matches in the 2014 tournament was estimated.

To verify the accuracy of prediction results for the match statistics model, values of the match statistics were placed in the model for each match. The model result was calculated and compared to the actual result for each match. The estimated response $\hat{y}$ was observed. If $\hat{y}$ was greater than 0, a predicted win for the team of interest was coded. If $\hat{y}$ was less than 0, a predicted loss for the team of interest was coded.

Results from the first to final rounds of the 2014 tournament were used to validate the match statistics model using differences in match statistics. The validation results for the

first round is given in Table 12. The model had an accuracy of 87.5%.

Table 12: Accuracy of least squares regression model developed by match statistics when validating first round of 2014

Point spread		Predicted		
		Win	Loss	Total
Actual	Win	13	2	15
	Loss	2	15	17
	Total	15	17	32
Overall Accuracy				87.5%

3.2 Validating second round using model developed

The validation results for second round using the match model are given in Table 13. The second round model had an accuracy of 93.75%.

Table 13: Accuracy of least squares regression model developed by match statistics when validating second round of 2014

Point spread		Predicted		
		Win	Loss	Total
Actual	Win	6	0	6
	Loss	1	9	10
	Total	7	9	16
Overall Accuracy				93.75%

3.3 Validating third and higher rounds using models developed

The validation results for third and higher rounds are given in Table 14. The model had an accuracy of 93.33%.

Table 14: Accuracy of least squares regression model developed using match statistics when validating third and higher rounds of 2014

Point spread		Predicted		
		Win	Loss	Total
Actual	Win	5	0	5
	Loss	1	9	10
	Total	6	9	15
Overall Accuracy				93.33%

3.4 Bracketing the 2016 tournament before tournament begins – Predicting

Since the match statistics will not be available before the tournament begins, differences of seasonal averages of the corresponding variables for both teams playing were collected and put into the match model to predict the winner of a volleyball match in the 2016 tournament. Results were predicted for each round by using the match model developed before the 2016 tournament begin by replacing differences between the match statistics of the two teams with differences in seasonal averages for the statistics from the two teams.

Differences of seasonal averages of significant variables were found for all teams playing in the first round and put into first round model. Differences of seasonal averages for each team predicted to play each other in the second round were then placed in the model and winners of this round were predicted. Differences of seasonal averages of variables found to be significant of teams predicted to play each other in the third round were placed in the model and winning teams predicted for this round. This process continued until a winner was selected.

The predicted results were then compared against the actual results for each round of matches for 2016.

Examples for each round in 2016 tournament

An example will be given as to how the match model for a particular round was used for each round in 2016 tournament.

The least squares regression model for first to final round developed by using differences in in-game statistics is:

$\hat{Y}$ = (0.04538* Diff_AttackK) + (8.12106 * Diff_Attack%)+ (0.21009* Diff_ ServeSA)

Nebraska played New Hampshire in the first round of the 2016 tournament. Data on differences of seasonal averages for significant variables were collected and displayed in Table 15.

Table 15: Nebraska and New Hampshire Statistics

Team	Games Won	Attack_K*	Attack_Percentage*	Serve_SA*
Nebraska	3	14.52	0.274	1.09
New Hampshire	0	11.82	0.198	1.73
Difference	3	2.7	0.076	-0.64

* Average per game for season

Using the model above, the match between Nebraska and New Hampshire had a predicted game difference of: $\hat{y}$ = (0.04538* 2.7) + (8.12106*0.076)+(0.21009* -0.64) = 0.61

Since $\hat{y}$ >0 this game was coded as a correctly predicted win for Nebraska, who won the match by a score of 3 to 0.

Round 1-6
Number correct: 30
Number incorrect: 33
Total: 63

Only the teams predicted to play each other in the second round were used in the model. The actual teams were not used all the time since predicting was done before the tournament started.

3.5 Results for Prediction by using models developed by match statistics

In 2016, a continuous process was used in verifying the models instead of doing round by round validations as in 2014. In other words, a complete bracket was filled out in 2016 before any match was played.

The match statistic model was used to predict the team in the first round who would advance to the next round. Once the teams in the second round were predicted, the same model was used to predict the winners of the second round. This process was continued for the third and higher rounds until the predicted final winner of the game was determined.

The prediction results for each round of 2016 tournament using the match model is given in Table 16.

Table 16: Prediction results of each round for 2016: (Least squares regression model developed by match statistics)

	Correct	Incorrect	Total games
First round	21	11	32
Second round	6	10	16
Third round	3	5	8
Fourth round	0	4	4
Fifth round	0	2	2
Final round	0	1	1
Overall Accuracy			47.62%

Accuracy of least squares regression model results is given in Table 16. The accuracy is only 47.62%. After the first round,

only the teams predicted to play each other in the second round were used in the model. The actual teams that played in the second round might not have been used. This was true for all matches starting in the second round and later. It was the teams that were predicted to play each other in each round that were used. All the predictions were filled out before the tournament started.

4. Conclusion

4.1 Validation - Models developed by using seasonal averages

The model developed by using differences of seasonal averages for the first round had approximately a 65.6% chance of correctly predicting the results. The model developed by using differences of seasonal averages in the second round had approximately a 68.8% chance of correctly predicting the results. The least squares regression model developed by using differences of seasonal averages for the third and higher rounds had approximately a 53.3% chance of correctly predicting the results.

4.2 Prediction - Models developed by using seasonal averages

In 2015, a continuous process was used to predict the winning team in each round before the tournament started instead of doing round by round predictions as in 2014. Namely, a complete bracket was filled out in 2015 before any game was played. When the differences of the seasonal averages for both teams for all the significant variables were considered for entry in the models developed by using differences of seasonal averages, the model had approximately a 63.5% chance of correctly predicting the winner of a volleyball game.

4.3 Validation - Models developed by using match statistics

To verify the accuracy of prediction results for the model developed by using match statistics, differences in the match statistics for both teams of all previously mentioned significant variables were placed in the model developed for the whole tournament. The model for the first round both had approximately a 87.5% chance of correctly predicting the results. The model for the second round had approximately a 93.8% chance of correctly predicting the results. The model for the third and higher rounds had approximately a 93.33% chance of correctly predicting the results.

4.4 Prediction - Models developed by using match statistics

When the differences of the seasonal averages for both teams for all significant variables were considered for entry in the regression model developed by using differences of match statistics and replacing these with differences in seasonal averages of the corresponding variables, the model had approximately a 47.6% chance of correctly predicting the winner of the tournament.

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

4.5 Overall comparisons

When predicting results for future tournaments when the match statistics are not known ahead of time and they need to be estimated, the models developed by using seasonal averages are better than the model developed by match statistics. Using differences of seasonal averages is better than using differences in ranks of seasonal averages.

5. References

1. NCAA Division I Women's Volleyball Championship. Retrieved October 10, 2017, from https://en.wikipedia.org/wiki/NCAA_Division_I_Women's_Volleyball_Championship
2. Road to the Championship. Retrieved October 10, 2017, from http://www.ncaa.com/championships/volleyball-women/d1/road-to-the-championship
3. Giatsis George. Statistical Analysis of Men's FIVB Beach Volleyball Team Performance. International Journal of Performance Analysis in Sport, 2008, 31-43.
4. Zhang D, Magel R, Degges R. Explaining and Estimating the Point Spread of Women's NCAA Volleyball Game. International Journal of Science and Research Methodology. 2018.
5. Ranking Summary. Retrieved October 10, 2016, from http://web1.ncaa.org/stats/StatsSrv/ranksummary
6. NCAA Official Volleyball Statistics Rules, 2017. Retrieved October 10, from http://fs.ncaa.org/Docs/stats/Stats_Manuals/VB/2017.pdf
7. Myers Raymond H. Classical and Modern Regression with Applications 2nd ed. Boston: PWS-KENT, 1990. ISBN 0-534-92178-7.