Regression Analysis of Division 1 Women's Volleyball

Ben Ellingworth

9/22/2022

Introduction

This paper explores the best team statistics for Division 1 Women's Volleyball to help predict a teams Win% (Win Percentage). Statistics from 325 D1 Volleyball teams were used in this study and tested to help find the best predictor. Five teams were removed from the study due to missing team statistics. Each stat was tested and compared to each other using a variety of methods. The study started with univariate regression models and also explored regression models involving two or more team statistics. The average Win% of the teams was .51 with a standard deviation of .24. The average amounts of sets for each team was 42.33 with a standard deviation of 6.38.

Table 1: Summary Statistics for Win PCT and Sets of D1 Volleyball Teams

Variable	Mean	Standard Deviation
Win PCT	0.51	0.24
Sets	42.33	6.38

Methods

R version 4.1.2 was used in this project. Data was collected and read onto R from the official NCAA Website. The data was updated as of Thursday, September 22.

To try and find the best predictor for win success, simple linear regression was used to help compare each variable. Each model was plotted with diagnostic plots, as well as scatter plots with a linear regression line. This helped provide a visual to check for correlation and also allowed to check for problems with each model. Transformation of both axes were considered but deemed unnecessary. During all of these tests, the response variable was Win%. However, in testing to see the variable that is the "best" predictor, the explanatory variable switched. Due to each team having a different number of games, each explanatory variable was divided by the number of sets the team has played to allow for equal representation. The explanatory variables that were used in this study include: Aces Per Set, Assists Per Set, Total Blocks Per Set, Digs Per Set, Kills Per Set, and Hit% (Hit Percentage = ((Kills - Errors)/Total Attacks)).

After testing each explanatory variable mentioned above, the coefficients of determination were stored in a table with the respective variable. 95% confidence intervals for the slope of the regressions were also collected. Intervals were stored in the same table as the coefficients of determination. A significance level of .05 was used for all hypothesis testing. Lastly, forest plots for each variable were created using each models regression coefficient and 95% confidence interval upper and lower bounds.

A similar process was used for multiple regression. The response variable stayed the same as Win%, but the explanatory variables used included: Hit% and Opponent Hit%, Kills Per Set and Assists Per Set, Total Blocks Per Set and Digs Per Set, and Aces Per Set with Opponent Hit%. Each of the variables were all fit into multiple regression models and Adjusted R-Squared was stored to determine how predictive the variables are. Adjusted R-Squared was used due to the error for complexity it takes into account.

Results

The best single variable predictor for winning in Division 1 Women's Volleyball is Hit%. As seen in Table 2 below, the R-Squared for Hit% was 0.66. This value is the largest out of all the different variables. The regression coefficient of 3.79 proves to be the largest rate of change between all models. This shows that there exists a correlation between Win% and Hit%. Plot 1 also shows a clear trend exists. The higher a teams Hit% is, the higher we can expect their Win% to be.

The confidence intervals for all the different variables are shown in Table 2. As seen in the table, we can say with 95% confidence, we expect an increase in Win% between (3.5,4.1) for each 1 unit increase of Hit%. The p value for the Hit% model was found to be <.001 which is well below the significance value of .05. This shows that there is a relationship between a team's Win% and Hit%.

Table 2: Single Variable Regression Results

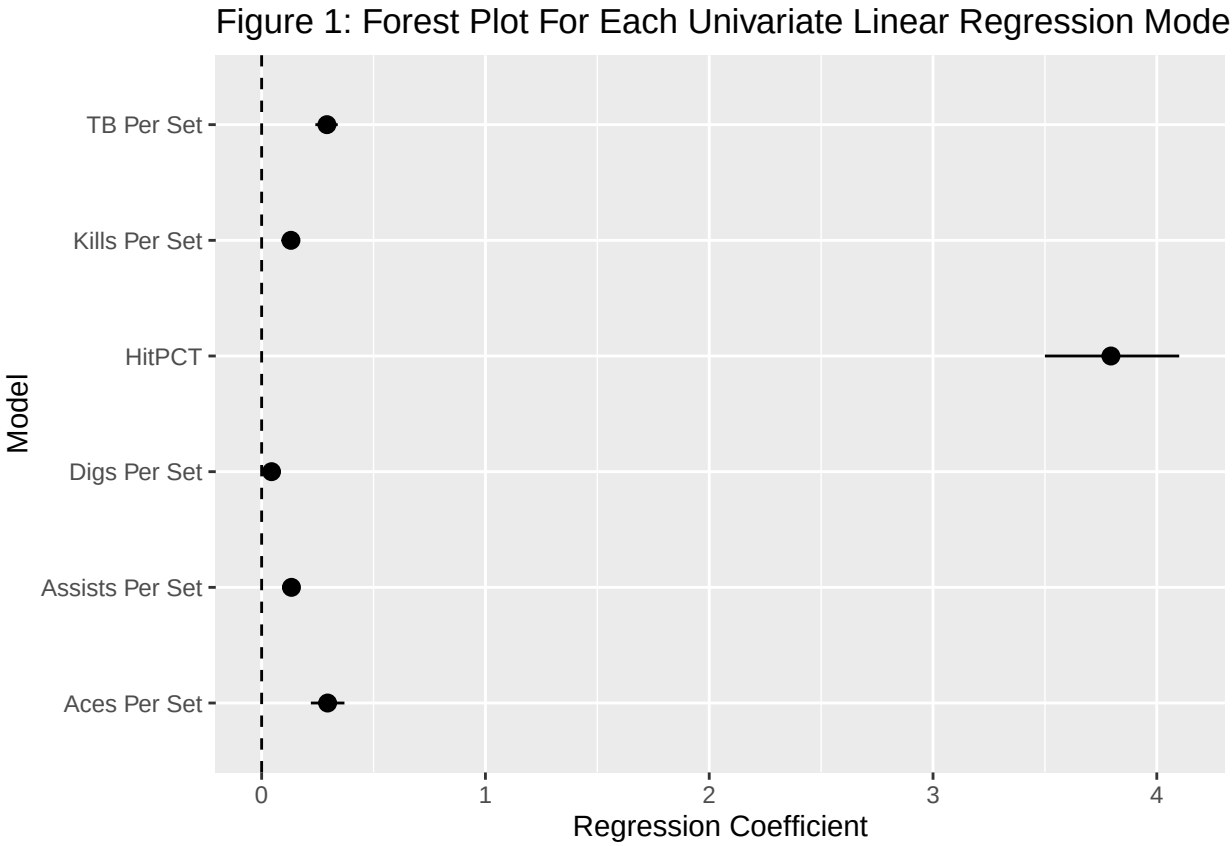
	Variable	R-Squared	95 CI Slope Lower Bound	Regression Coeffecient	95 CI Slope Upper Bound
1	Hit PCT	0.66	3.50	3.79	4.10
6	Kills Per Set	0.56	0.12	0.13	0.14
3	Assists Per Set	0.53	0.12	0.13	0.15
4	Total Blocks Per Set	0.28	0.24	0.29	0.34
2	Aces Per Set	0.16	0.22	0.29	0.37
5	Digs Per Set	0.09	0.03	0.04	0.06

Note:
CI = Confidence Interval

When working with real world data, not every aspect can fit perfectly into a linear model. Diagnostic plots were taken for the purpose to try and point out errors with each model. Roughly two high leverage points were found when using Hit% as the explanatory variable. This helps highlight that the results are

meaningful and can be very useful, but still have limitations.

Figure 1 shows the forest plot that includes each simple linear model used. This plot is useful to show that the Hit% regression coefficient and interval are well above the other models. As a result, we can conclude that there is a greater rate of change involving Hit% than other models. This illustrates the importance of Hit% when trying to predict Win%.



Using multiple regression which accounts for multiple variables, the most predictive team stats together were a team's Hit% and their Opponents Hit%. So, while it is very important to focus on increasing a team's Hit%, it is also important to try and limit the opponents. As seen below in Table 3, the Adjusted R-Squared for the model incorporating both Hit%'s is 0.76. This is an increase from the singe variable regression and well above all other combinations tested.

Table 3: Multiple Variable Regression Results

Variables	Adjusted R-Squared
Hit PCT and Opponent Hit PCT	0.76
Kills Per Set and Assists Per Set	0.56
Aces Per Set and Opponent Hit PCT	0.41
Total Blocks Per Set and Digs Per Set	0.36