An Analysis of NCAA Women's Volleyball Scoring Kills with Assists*

Math 261A Project 1

Robert Yav

October 29, 2025

The purpose of this paper is to analyze the relationship between a woman's volleyball team's assists and kills over the course of a season. We explore data on the 2024-2025 NCAA Women's Volleyball season's team's points scored and the number of assists per team by performing a simple linear regression on the kills as explained by the number of assists. It found that there was a positive relationship between a team's overall number of kills versus the team's total assists over the season. This study can provide exploratory analysis and insight to other investigators or teams looking to model and predict future team outcomes using recent volleyball statistics.

Introduction

Volleyball is a sport that requires elite athleticism, hand-eye coordination, open communication, and most importantly, the ability to combine all of these aspects into a functioning team. One fundamental concept of volleyball is the three touch play: Receive, set, and spike. There are, however, situations in which the ball will not need an additional touch, or assist, to get a point. A kill is a situation in which a team makes a play to render their opponents unable to prevent a bounce on the court on their playing area. This does not include situations in which the opponent fails to bring the ball from their side to their opponents' side, which would result in a point for their opponent. An assist is counted when a set, pass, or dig is made that directly allows a teammate to make a kill during a set. Then begs the question: Does one need to make assists to make more kills in volleyball? This project examines the relationship between the number of kills made for a team to the number of assists a team makes in a volleyball season. We expect there to be a positive This investigation may provide insight to coaches and players who want to improve their team's point totals by identifying what to

*Project repository available at: [https://github.com/rubertyao/NCAAWomensVolleyballAnalysis] prioritize improving during training and in game. Previous researchers in this field attempted to categorize player and team skills into a scale-based system which split the assist category we are interested in into two categories (Serve Reception and Dig) (Eom and Schutz 1992), however for this experiment we were more interested in particularly those two categories as one metric. In the next sections, we perform a simple linear regression on kills as explained by assists, paying attention to any irregularities that may occur when we expect a positive linear relationship between those two variables.

Data

The data set used was collected from the Division 1 Women's NCAA Volleyball 2022-2023 season ("Team Statistics for Division i Women's Volleyball – SCORE Sports Data Repository" 2025). Division 1 is the highest division in college sport divisions, which means that the highest caliber of skill at the college level is displayed in these data. There are 344 rows and 14 variables, with each row representing a team at the Division 1 level from the 2022-2023 season. The 14 variables include: Team, Conference, region, aces per set, assists per set, team attacks per set, blocks per set, digs per set, hitting percentage, kills per set, opp hitting percentage, W, L, and win loss percentage. My analysis focuses on two of the variables, kills per set and assists per set.

assists per set - The average amount of sets, passes, or digs to a teammate that directly result in a kill per set. Digs are usually made immediately after the opposition attempts a kill to prevent a ball from hitting the ground and pass is made usually towards a setter to "set up" a ball for a hitter to attempt a kill on the opposing team.

kills per set - Average amount of hits that directly result in a point per set.

A quick observation of Figure 1 indicates two potential outliers and a noticeable positive correlation between the variables. The outliers, however, under closer inspection, cannot be reasonably removed, as the teams performed lower compared to the other teams in the season (i.e. 28 losses to 0 wins).

Methods

We fit the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

β_1 represents the expected point change in kills per set per point change in assists per set.

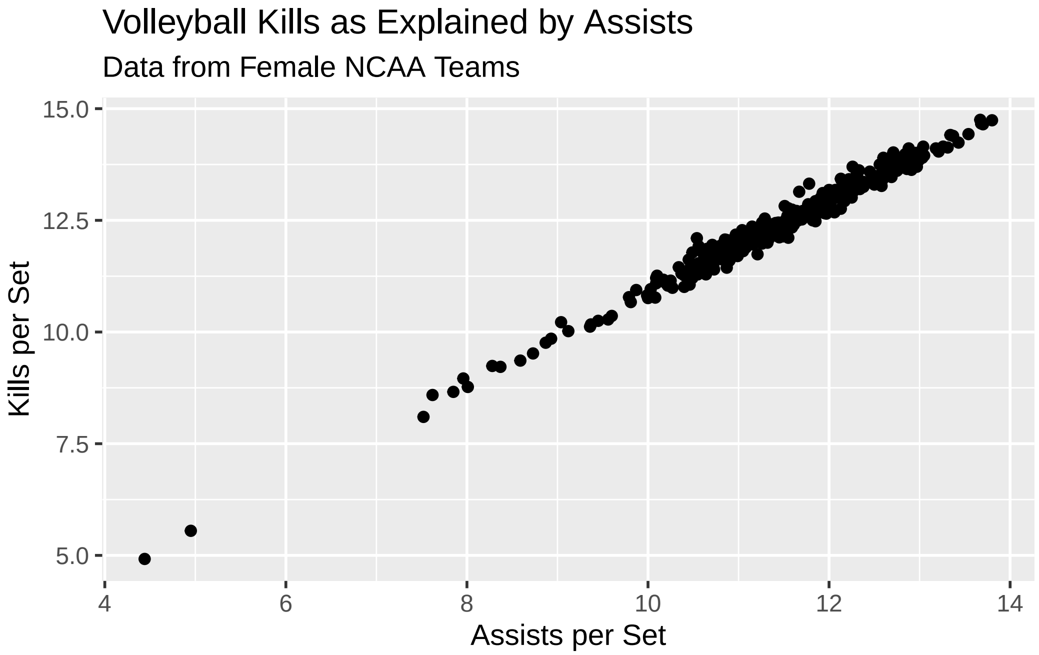

Figure 1: Scatterplot of Volleyball Kills per Set and Volleyball Assists per Set

β_0 represents the intercept or expected number of kills per set if we had a team with 0 assists per set. In a sports organization with players playing at the highest caliber, players will be making assists in a game, so this statistic holds no practical relevance.

ϵ_i represents independent, uncorrelated, normally distributed error terms with mean 0 and constant variance σ^2

X_i represents the i^{th} team's assists per set as a predictor in the model.

Y_i represents the i^{th} team's kills per set as a response in the model.

We implement the analysis using R (R Core Team 2025). Visual aid was done in ggplot2 (Wickham 2016). Tables were done with aid from knitr (Xie 2025).

Estimate	Standard Error	t-value	P-Value
0.595640	0.0881799	6.754829	0
1.029573	0.0076719	134.201136	0

Figure 2: Summary Table of Volleyball Model

We attempt a two-tailed hypothesis test for β_1 with $\alpha = 0.05$ in Figure 2:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

The slope parameter $b_1 = 1.023$, meaning that the model estimates that for each point change in assists, the expected point total should change by 1.023. The intercept parameter $b_0 = 0.596$, meaning that for 0 assists, the expected amount of points a team would have would be 0.596.

Since our p value in Figure 2 P(t<134.201)= 2.642e-291 is much smaller than our $\alpha = 0.05$, we can reject the null hypothesis, meaning that there is statistically significant evidence of a relationship between assists and kills. Note that in Figure 2 we have obtained a value of 0, meaning that the value is so low that R cannot reasonably show the number in its entirety. We note that since the number is so low, the chances of a true linear relationship between the two variables is so high that a hypothesis test is almost not needed to confirm it. $R^2 = .9819$, meaning that our model explains about 98.19% of variation in our sample. This is a good indicator that our model is performing well.

In order for our linear model to hold, we make certain assumptions: our data must be valid for our research question, our sample needs to adequately represent our population of interest, we require a linear relationship to make sense between our predictor and response variable, independence of errors, equal variance of errors, and normality of errors, which we will address with a plot between our x and y, scatterplot of residuals versus fitted values, and a normal qqplot.

Validity and representativeness of our population of interest of our data is satisfied by design of our experiment, since it was built around our data. Independence of the errors is satisfied by design of the data, since each team appears only once and is recorded independently of the others.

A violation of linearity would mean that fitting a simple linear regression would not make sense for our data, and that we would need to attempt to transform or consider other models. Observing the scatterplot in Figure 3 between volleyball kills per set and assists per set confirms that a linear relationship between our predictor and response variable would be valid. The red line represents the expected value of kills per set for a team with a given number of assists per set.

A violation of the assumption of equal variances would make the hypothesis test and its associated p-value unreliable, as OLS becomes inefficient with our parameter estimates. An observation of the residuals versus the fitted values in Figure 4 revealed that our assumption of equal variance has been upheld.

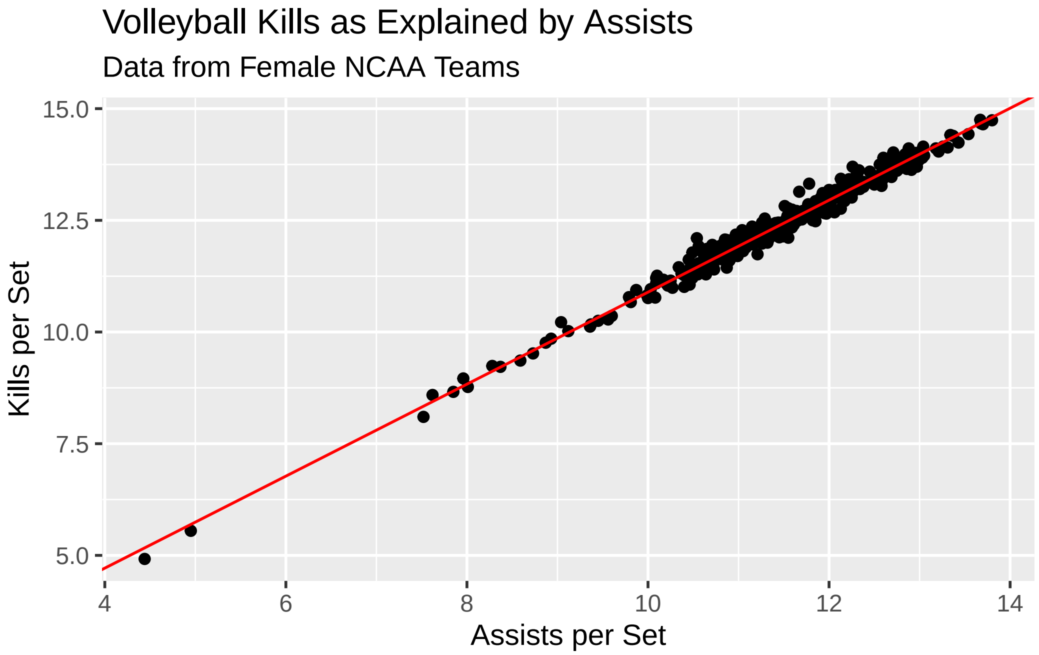

Figure 3: Scatterplot of Volleyball Kills per Set and Volleyball Assists per Set with Fitted Linear Regression Model

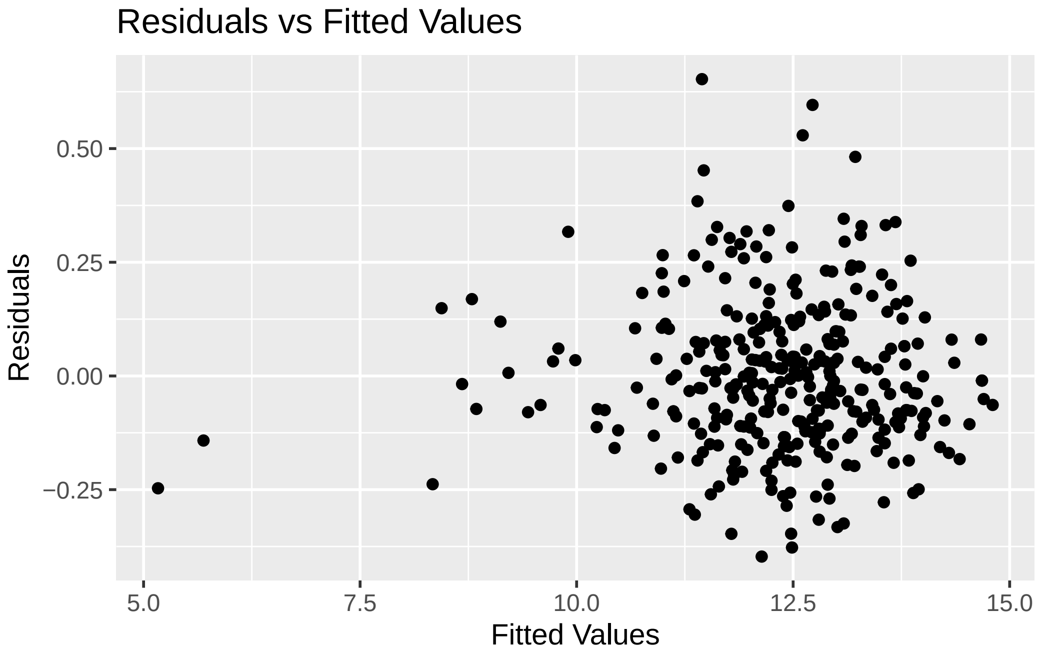

Figure 4: Residuals vs Fitted Values Plot

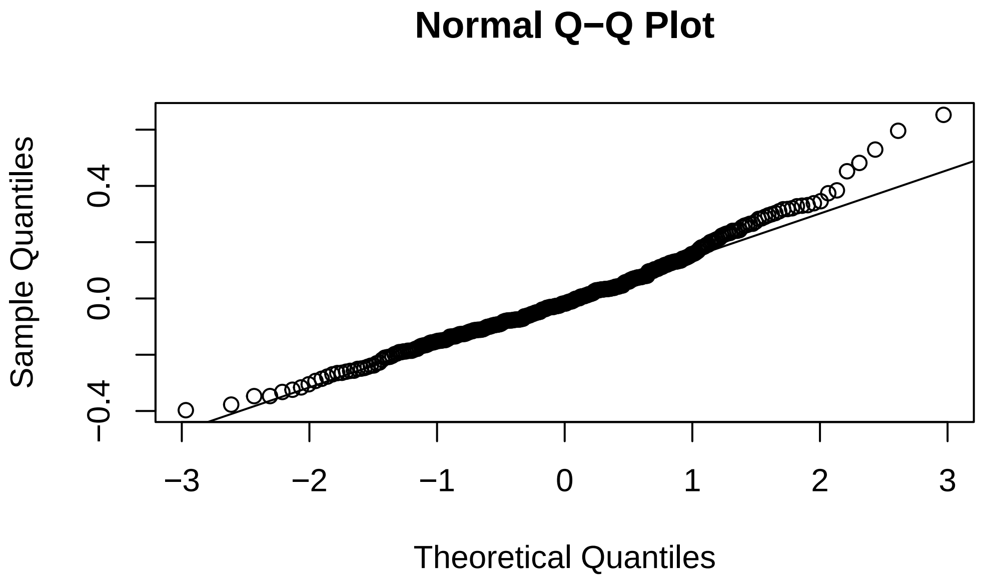

Figure 5: QQ-Plot of Residuals

A major violation of normality of errors would bias the parameter estimates, providing an unreliable p-value. A quantile-quantile plot in Figure 5 comparing the distribution of our values to a normal distribution revealed a large tail at the right end, signifying that there were more values on the extreme than expected under a theoretical normal distribution, represented by the line drawn. This violates the assumption of normally distributed errors, but since our sample size is sufficient enough to overlook this, we can safely continue.

Results

We have the results of our two-tailed t-test on our parameter β_1. In order There is a positive correlation between the number of assists a team makes and the number of kills a team gets. In our model, each additional assist is expected to land us an additional $b_1 = 1.023$ kills. We take caution in this conclusion though, as an assumption that allows us to make an estimate on the true value of β_1 may have been broken, normality of errors. Since we have a high sample size (344), this violation may be overlooked.

Discussion

We are aware that this data only covers D1 women's NCAA athletes, which excludes all men, recreational, a co-ed players, as well as any lower division women's teams. Also, we can only make statistical inferences on D1 women's teams, since the data does not magnify into individual players.

Our assumption of normality seems to have been broken by our quantile-quantile plot. This is potentially due to multiple factors, such as the idea that our values are mostly on the extreme right end of the data, which may have to do with the idea that many of these teams play at a high level, so higher values of assists and kills may have a higher probability of appearing than lower values. This violation is not a large departure from normality, and we have a large sample size (344), therefore we can argue that this departure does not create any serious problems for our model.

The fact that a kill must be present for an assist to be counted in this dataset may serve as evidence to an obvious correlation between the two variables, so future analyses should investigate "passes and sets that do not result in kills."

References

Eom, Han, and Robert Schutz. 1992. "Statistical Analyses of Volleyball Team Performance." *Research Quarterly for Exercise and Sport* 63 (April): 11–18. https://doi.org/10.1080/02701367.1992.10607551.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

"Team Statistics for Division i Women's Volleyball – SCORE Sports Data Repository." 2025. NCAA. 2025. https://data.scorenetwork.org/volleyball/volleyball_ncaa_team_stats.html.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Xie, Yihui. 2025. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.