An Analysis on NCAA Woman's Volleyball Statistics[*]

Math 261A Project 1

Robert Yav

October 7, 2025

This paper explores the relationship between the NCAA Woman's Volleyball teams' points scored to 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 although there is a positive correlation between the two variables, many assumptions of a regression on the variables have been violated and therefore require additional amendments to the model fitted.

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 to get a point. 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. This investigation may provide insight to coaches and players who want to improve their team's point totals by identifying what to prioritize improving during training and in game.

Data

The data set used was collected from the Division 1 Woman's NCAA Volleyball 2022-2023 season ("Team Statistics for Division i Women's Volleyball – SCORE Sports Data Repository" n.d.). Division 1 is the highest division in college sport divisions, which means that

[*] Project repository available at: [https://github.com/rubertyao/NCAAWomansVolleyballAnalysis]

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_pctg, kills_per_set, opp_hitting_pctg, W, L, and win_loss_pctg. 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

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

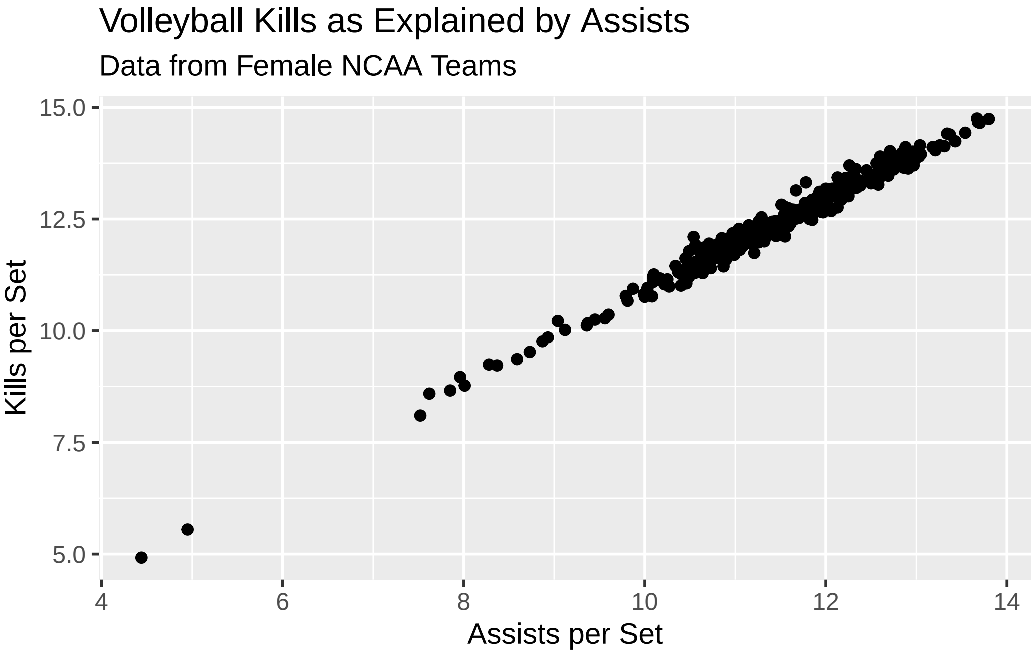

Figure 1: Scatterplot of Volleyball Kills per Set and Volleyball Assists per Set

A quick observation indicates two potential outliers and a noticeable positive pattern 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, which still is important to include. (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.

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.

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

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

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.595640	0.0881799	6.754829	0
vb_data$assists_per_set	1.029573	0.0076719	134.201136	0

Summary Table of Volleyball Model

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

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

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 P(t<134.201)$\sim$ 0 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. $R^2 = .9819$, meaning that our model explains about 98.19% of variation in our sample. This is a good indicator that our model is performing well.

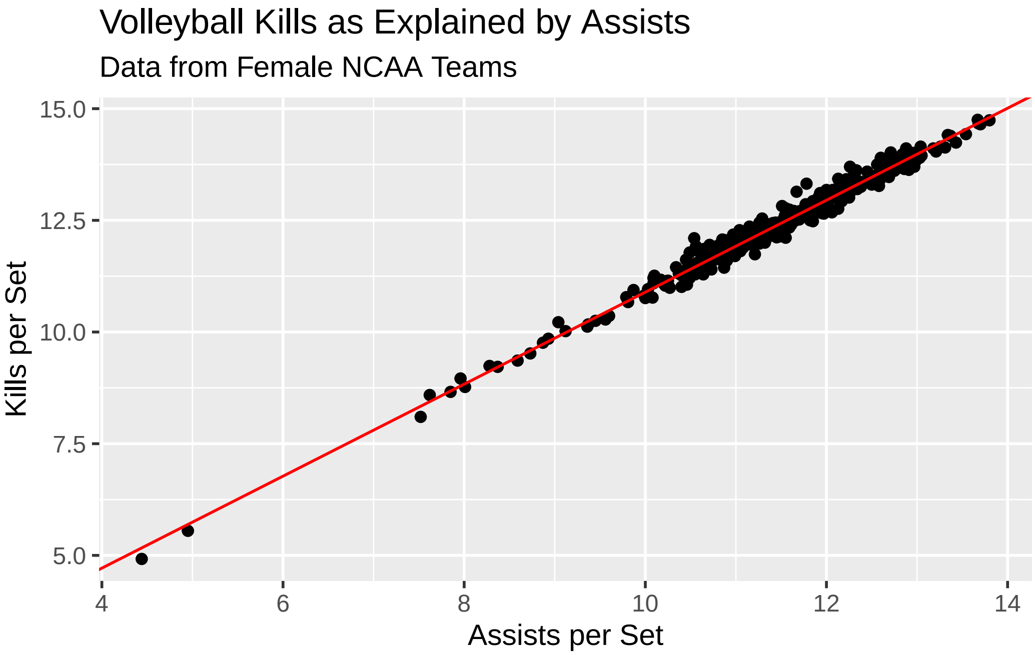

Figure 2: Scatterplot of Volleyball Kills per Set and Volleyball Assists per Set with Fitted Linear Regression Model

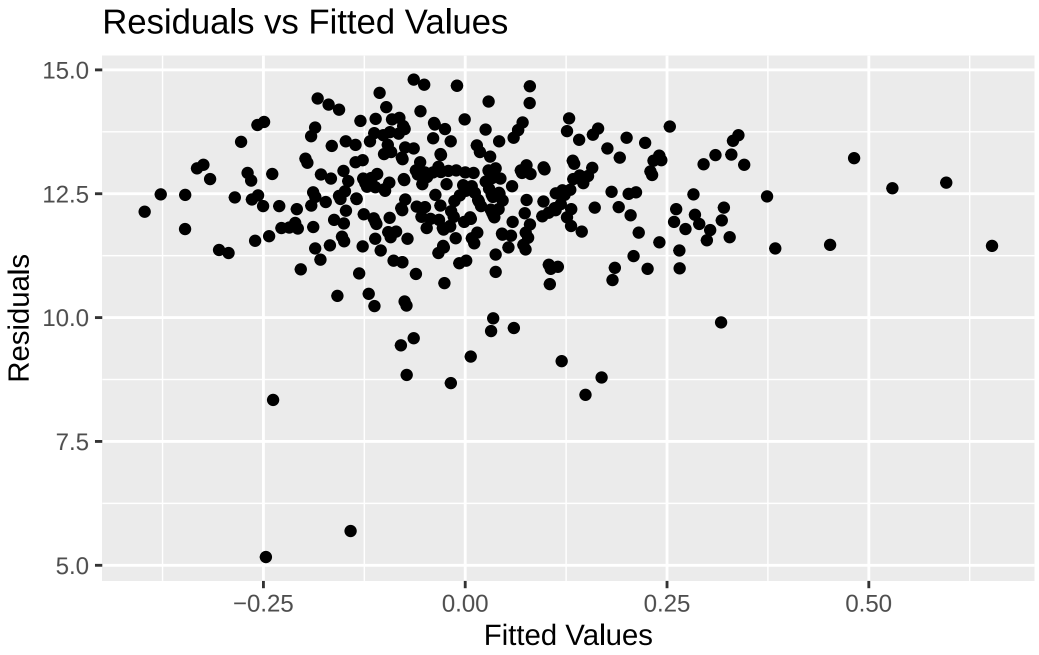

Figure 3: Residuals vs Fitted Values Plot

An observation of the residuals versus the fitted values revealed that our assumption of equal variance have been broken. As we can see, most of our residuals are positive, meaning that our model generally underestimates the true number of points scored, as the observed values are generally larger than the predicted values ($Y_i > \hat{Y}_i$).

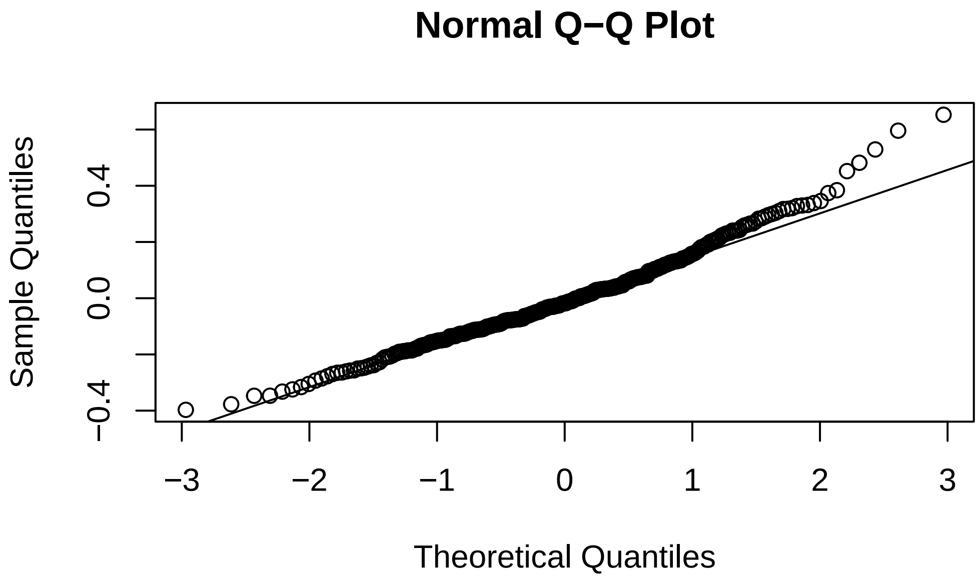

Figure 4: QQ-Plot of Residuals

A quantile-quantile plot of the residuals revealed a large tail at the right end, signifying that there were more values on the extreme than expected under a normal distribution. This violates the assumption of normally distributed errors, but with a high enough sample size, this may be overlooked.

Results

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 the assumptions that allow us to make an estimate on the true value of β_1 have been broken, constant variance and normality of errors. Although this model does have a good fit due to its high $R^2 = .9819$ and low p-value $p \sim 0 < 0.05$, again due to violations of normality and constant variance, we may require additional variables, data transformation, or fitting a different model besides simple linear regression.

Discussion

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

Our assumptions of normality and constant variance have been broken by our residual to fitted values 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, denoting that most of these D1-level teams seek out high-scoring athletes and have maintained D1 status due to this requirement. The fact that in order for an assist to be made, a kill must occur could have an influence on the violation of our assumptions, so better variable selection to also include "passes and sets that do not result in kills" in "assists" could be investigated.

References

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." n.d. NCAA. Accessed September 23, 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/.