**Introduction**

For this activity, you will be exploring data from the 2022-2023 NCAA Women’s Volleyball season. Specifically using team match statistics such as kills and assists, you will fit single linear and multiple linear regression models to predict their season long win percentage. Additionally, you will investigate the meaning of the different variables and analyze the relationships they have with each other. Discovering which variables make the most effective predictors, you will gain an insight for team metrics that produce the most wins.

**Learning Goals**

This activity will help you understand:

* Understand the concept of multicollinearity and its implications in multiple linear regression.
* Identify and interpret correlations between variables.
* Analyze the effectiveness of different predictors in predicting win percentage.
* Calculate and interpret Variance Inflation Factor (VIF) to assess multicollinearity.

**Data**

The data used in this activity was scraped from the NCAA website where you can find a multitude of variables relating to the 2022-2023 Division 1 Women’s Volleyball season. Containing an observation for each team, there are 334 total rows and 15 columns in the data set. The full data (volleyball\_2022\_23.csv) is available on the GitHub repo associated with this module.

**Methods**

For this activity, you will need to produce scatterplots, identify correlation between variables, create linear models to examine their effectiveness, and calculate VIF. For R users, you will need the following packages:

* tidyverse
* car
* readr

**Exercises**

Model 1: Kills

Identify the correlation between kills\_per\_set and win\_loss\_pctg.

r = 0.79

Fit a model that uses kills\_per\_set to predict win\_loss\_pctg. Is there evidence that kills\_per\_set is a useful predictor?

Yes, there is strong evidence that kills\_per\_set is a useful predictor for win\_loss\_pctg (p-value ≈ 0.000).

Model 2: Assists

Identify the correlation between assists\_per\_set and win\_loss\_percentage.

r = 0.78

Fit a model that uses assists\_per\_set to predict win\_loss\_pctg. Is there evidence that assists\_per\_set is a useful predictor?

Yes, there is strong evidence that assists\_per\_set is a useful predictor for win\_loss\_pctg (p-value ≈ 0.000).

Model 3: Kills + Assists

Fit a model with assists\_per\_set AND kills\_per\_set predicting win\_loss\_pctg. Are both predictors still significant? What seems to have happened with the slope coefficient of assists\_per\_set? Did the p-values change?

The predictor assists\_per\_set is no longer significant. The coefficient for assists\_per\_set is now negative although before it indicated an increase in win\_loss\_pctg. This model that uses both assists\_per\_set and kills\_per\_set has an Adjusted R-squared value that is barely larger than that of the previous models despite both predictors being strong in their respective models.

Reasoning

Create a scatterplot of kills\_per\_set against assists\_per set. What do you notice?

A graph with a dotted line

Description automatically generated

The relationship between the two variables displays a strong linear relationship with very little variance.

What is the correlation between kills\_per\_set and assists\_per set?

r = 0.99

In volleyball, a kill is awarded to a player any time their attack is unreturnable by the opposition because it is the direct cause of the opponent not returning the ball. An assist is awarded when a set, pass, or dig to a teammate results in that teammate attacking the ball for a kill.

Given these definitions, why do you think we lose significance when we use both kills and assists to predict win\_loss\_pctg? Does the addition of assists\_per\_set improve our model?

We lose significance because in order to have an assist, there must always be a kill. Kills can occur without an assist, but it is an uncommon event as can be seen in the scatterplot and by the correlation. Therefore, the two variables are generally equal in predicting win\_loss\_pctg so the addition of assists\_per\_set does not improve our model in any way. Though the R-squared value improves, it is no longer a good indication of model quality as one of the predictors is not significant.

Which of the three models do you think is most effective?

The model using kills\_per\_set is the most effective as the single predictor has strong significance and the R-squared value is higher than that of the model using assists\_per\_set. Naturally, this makes sense given the defintion of kills and assists in volleyball.

Using the model that you decided on, provide an interpretation of the slope coefficient in context.

On average, an increase of 1 kill per set is associated with a 0.127 increase in win percentage.

Using the model that you decided on, provide an interpretation of the intercept coefficient in context.

When kills per set is equal to 0, we expect that there is a win percentage of -1.065.

Multicollinearity in multiple linear regression occurs when explanatory variables are highly correlated. It can result in unstable coefficient estimates and hinder interpretations of predictor significance. Removing these variables can address multicollinearity and improve the model's stability and interpretability.

One method of identifying multicollinearity is through calculating VIF. The calculation determines the strength of the correlation between each pairing of explanatory variables. VIF values greater than 5 are generally considered to imply an issue with multicollinearity

Using the equation above, Calculate the VIF of assists\_per\_set and kills\_per\_set below. Does the result confirm the presence of multicollinearity in the model?

= 55.25

= 55.25