Jake Meagher
Stat 98: Module 1

Correlation and Linear Relationships

The collection and interpretation of numerical data can reasonably be construed as the foundation of statistics. Statisticians often take multiple data sets and compare them in order to look for the presence of some sort of relationship between variables, which are sets of measured quantitative values. The utilization of correlation is a technique that allows statisticians to do just that.

Specifically, correlation measures whether or not there is a linear relationship between two variables. A relationship is said to be linear if the data points which come from the two variables would lie on an approximately straight line when placed on a scatterplot. Assuming the two variables being compared are linearly related, correlation also measures the strength of that linear relationship. The strength of the relationship can be represented by what is known as a correlation coefficient. The correlation coefficient has no units, which enables statisticians to compare linear relationships across different samples, even if some present different variables.

The correlation coefficient can assume any value between -1 and 1. A value of 0, while within the range of possible values, represents the lack of a linear relationship between the two variables. Meanwhile, a value of -1 or 1 represents the presence of a perfect linear relationship between the two variables. A perfect linear relationship suggests that given the value of one variable, you can determine the exact value of the second variable. Under these conditions, if you were to place the two variables on a scatterplot, a straight line could be drawn through all the points. Correlation coefficients of -1, 1, and 0 are rare in practice, however. Typically, the correlation coefficient takes some non-integer value between -1 and 1. The closer the value of the correlation coefficient is to -1 or 1, the stronger the linear relationship is. Likewise, the closer the correlation coefficient is to 0, the weaker the linear relationship is.

While the value of the correlation coefficient determines the strength of a linear relationship, the sign in front of the correlation coefficient determines the direction of the relationship. When a correlation coefficient is positive, the two variables are said to have a positive relationship, which means as the values of one variable increase, the values of the other variable are expected to increase. Meanwhile, when a correlation coefficient is negative, the two variables are said to have a negative relationship, which means as the values of one variable increase, the values of the other variable are expected to decrease.

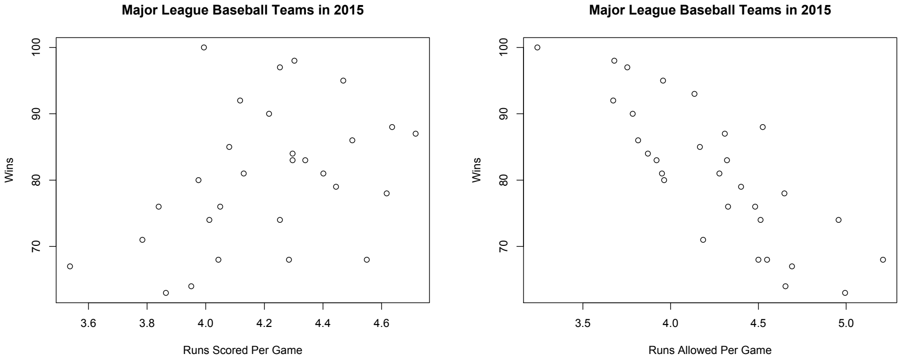

Figure 1: Based on data collected from the 2015 Major League Baseball season, the correlation between Runs Scored Per Game and Wins (left) is 0.39, which indicates a positive but relatively weak linear relationship between the two variables. The correlation between Runs Allowed Per Game and Wins (right) is -0.81, which indicates a negative and relatively strong linear relationship between the two variables.

The difference between positive and negative relationships is illustrated in the scatterplots above, which present data from the 2015 Major League Baseball season. In baseball, when you score more runs, you tend to win more games, and when you allow more runs, you tend to win fewer games. Thus, the scatterplot on the left depicts a positive relationship, while the scatterplot on the right depicts a negative one. These plots also provide information about the strength of these two relationships. The points in the right-most plot look like they would lie on a relatively straight line, while the points in the left-most plot appear far more scattered, indicating that the linear relationship between Wins and Runs Allowed Per Game is the stronger of the two relationships. The correlations listed above support this claim, leading us to believe that it is more important to stop runs than to score them in order to win games. The unitless nature of the correlation coefficient makes this comparison, like many others, possible, proving why correlation is such an integral part of statistics.