

Correlation and Linear Regression Statistical Study

Introduction

As a devoted basketball fan, I've watched NBA games on TV, online and even once in an arena. Since I've started watching basketball, I've never had a favorite team, but I do have several favorite players on different teams. Here's a few of my favorite players; From the Brooklyn Nets Kyrie Irving , Milwaukee Bucks Giannis Antetokounmpo, Houston Rockets James Harden, and my personal favorite dynamic duo the splash brothers from the Golden State Warriors Stephen Curry and Klay Thompson. Unlike the short NFL season which is only sixteen games and unlike the long MLB season that has 162 games during a single season. I feel as if the NBA season has just the right amount of games. The basketball teams in the NBA only play 82 regular season games. As a long time, basketball fan, I noticed that as players are drafted into the league as rookies, they get playtime but as they gain more experience and as they develop, they get more playtime, but does that mean their average points per game also increases. Therefore, in this paper I will

utilize common basketball statistics and explore the connection between the average minutes per game and the average points per game of fifty individual players.

In this project of correlation and linear regression statistical study we are trying to determine the relationship between two variables; the independent variable which is x and the dependent or otherwise known as response variable known as y . We want to determine how the different values of the independent variable correlate with the response variable.

The Variables:

X: Average Minutes played per game

Y : Average Points scored by a player per game

Data Collection:

In order to collect the appropriate data for this study I used the official website of the NBA.

[https://stats.nba.com/leaders/?Season=2018-19&SeasonType=Regular
Season](https://stats.nba.com/leaders/?Season=2018-19&SeasonType=RegularSeason)

My hypothesis is that the response variable will have a positively skewed distribution. There will also be a strong positive correlation. I hypothesize this because if the player receives more play time that would mean they would have a higher chance of scoring more points.

Analysis

During the data collection process I gathered data from *NBA.com* since all the statistics that I needed were displayed on the website. I collected the average minutes played per game and the average points scored from fifty players. Then I organized the data into two columns on excel. As mentioned previously the X variable would be the average minutes player game and the Y variable would be the average points scored by a player per game. For the first part of this analysis I collected the five-number summary of the dependent variable (Y) using the values that I had collected. I was able to find that the maximum was 36.1, the minimum was 16.6, the median was 21.05. Moving on to the 1st and 3rd quartiles I was able to find that the first quartile otherwise defined as the middle number between the smallest number and the median of the data

set was equivalent to 18.175. Moreover, the third quartile otherwise defined as the middle number of the part of data which is greater than the median which was equivalent to 24.425. Using the data, I was able to calculate my x-mean which is essentially the average of all fifty x values. Which equaled to 32.95 after calculating the x-mean I calculated the y-mean which equaled to 21.57. From this data we can tell that the average minutes played within all fifty players is 32.95 and the average points scored within all fifty players is 21.57. Then moving on to calculating s_x and s_y otherwise known as the standard deviation for both x and y values. After calculating the standard deviation for both x and y. The calculation for x equaled to 2.29 and for y it is 4.03. As for the correlation coefficient otherwise known as (r) after calculations it equaled to 0.57. Moreover, using the y-intercept, slope and values from X I was able to calculate the $\hat{y}$. And to find the residuals it was simple subtraction from the Y values and the $\hat{y}$ values. Moving over to the construction of the scatterplot I used the values of X and Y in order to find the regression line that include a y-intercept of 11.7 and a slope of 1.01. Now to further discuss the data we're going to discuss the

skewness of the histograms. In statistics skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined. But in this study the histograms displays a left skew which shows that the distribution is positively skewed. Now moving on to the scatterplot I would identify it as a weak positive correlation. To compare that to my predictions I predicted a positively skewed histogram and a strong positive correlation. Referring to the data I would say overall my predictions agreed with the data.

Conclusion

To conclude my findings, I discovered that there was a weak positive correlation between the average minutes and the average points. A weak positive correlation would indicate that while both variables tend to go up in response to one another, the relationship is not very strong. In real life this would mean that the average minutes a player plays per game does correlate with the average points they can score however the relationship between the two is very weak. This means a rookie

shouldn't worry about his minutes he just needs to focus on performing to the best of his ability.