Correlation & Regression

Discussion 11

Data C8: Foundations of Data Science — UC Berkeley, Fall 2025

An important aspect of data science is using data to make *predictions* about the future based on the information that we currently have. A question one might ask would be, “Given the amount of time a student studied for an exam, what would we predict their grade to be?” In order to answer this question, we will investigate a method of using one variable to predict another by looking at the *correlation* between two variables.

1. Standard Units and Correlation

> **The Correlation Coefficient:** The average of the product of x and y when both are in standard units.
>
> This coefficient is a number between -1 and 1 that measures the strength and direction of the *linear* relationship between two variables, x and y.

(a) When calculating the correlation coefficient, why do we convert data to standard units?

(b) Write a function called `convert` which takes in an array of elements called `xs` and returns an array of the values represented in standard units.

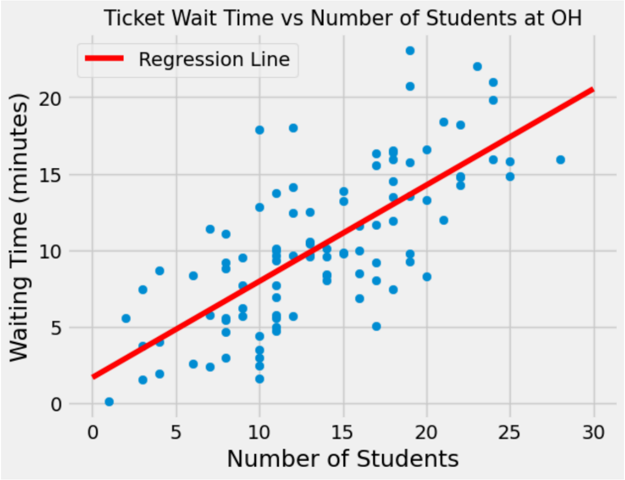
```
def convert(xs):

    sd = _________________________________

    mean = _______________________________

    return _______________________________
```

(c) Write a function called `correlation` which takes in a table of data `tbl` containing the column names `x` and `y` and returns the correlation coefficient.

```
def correlation(tbl, x, y):

    x_su = ____________________________________________________________________________________

    y_su = ____________________________________________________________________________________

    return ____________________________________________________________________________________
```

3. Linear Regressi(OH)n

You just submitted a ticket at Office Hours and would like to know how long it will take to receive help. However, you don't believe the estimated wait time displayed on the queue to be very accurate, so you decide to make your own predictions based on the total number of students present at OH when you submitted your ticket. You obtain data for 100 wait times and plot them below, also fitting a regression line to the data.



(a) Suppose that you submit a ticket at Office Hours when there were a total of 20 students present. Based on the regression line, what would you predict the waiting time to be?

(b) You go to Office Hours right before a homework assignment is due, and despite safety concerns, you observe 70 students at Office Hours. Would it be appropriate to use your regression line to predict the waiting time? Explain.

(c) When constructing your regression line, you find the correlation coefficient r to be roughly 0.73. Does this value of r suggest that an increase in the number of students at Office Hours *causes* an increase in the waiting time? Explain.

(d) Suppose you never generated the scatter plot at the beginning of this section. Knowing ***only*** that the value of r is roughly 0.73, can you assume that the two variables have a linear association?

○ A. Yes, r tells us the strength of a linear association and a high value of r always proves that the two variables have a linear association.

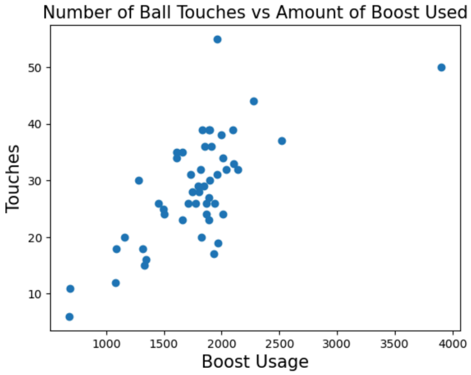
○ B. Yes, because if we can compute the value of r, the two variables must have a linear association.

○ C. No. A high value of r does not necessarily imply that the relationship between the variables is linear.

○ D. No, the value of $r = 0.73$ is not high enough to imply a linear association.

4. This is Regression!

Conan has an unhealthy addiction to *Rocket League*, a game where players play soccer but with cars instead of people. Players can pick up boost pads that are scattered across the field, which players can use to make their cars go faster! Conan plays 50 games and records how much boost he used, as well as how many times he touched the ball in a given game.



(a) Select the correct option that corresponds to the data.

○ There appears to be a positive association.

○ There appears to be a negative association.

○ There is not enough information.

(b) Select the correct option that corresponds to the data.

○ An increase in Boost Usage *causes* an increase in Touches.

○ An increase in Boost Usage *does not cause* an increase in Touches.

○ There is not enough information.

(c) Conan runs some calculations and obtains the following statistics:

- The **correlation coefficient** between Touches and Boost Usage was approximately **0.705**.
- The average number of Touches was 28.54 with a standard deviation of 9.51.
- The average of Boost Usage was 1773.4 with a standard deviation of 471.7.

i. Conan touched the ball 40 times in one of his games. What is this in standard units?

ii. Conan wishes to fit a regression line to the data. What would be the slope and intercept of the regression line in original units?

iii. What would be the slope and intercept be if the data were in standard units?