# Lab 4b: More Correlation

Stat 131A, Fall 2018

**Learning Objectives:**

- Use a correlation coefficient to describe the direction and strength of a linear relationship.
- Distinguish between association and causation.
- Identify lurking variables that may explain an observed relationship.

## General Instructions

- Write your solutions in an `Rmd` (R markdown) file.
- Name this file as `lab04b-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `lab04b-gaston-sanchez.Rmd`).
- Knit your `Rmd` file as an html document (default option).
- Submit your `Rmd` and `html` files to bCourses, in the corresponding lab assignment.

---

## Problem 1

Indicate whether the following statements are True or False.

a. A horizontal line has no slope.

b. $r$ reflects the slope of the scatterplot.

c. If a line has a positive slope, y-values on the line decrease as the x-values increase.

d. The magnitude of $r$ indicates the strength of the linear relationship.

e. $r$ ranges from 0 to 1.

f. A value of $r$ close to 0 suggests that the variables have a strong linear relationship.

g. The sign of $r$ and the sign of the slope of the regression line are identical.

h. A perfect positive correlation $r = 1$ indicates that one variable causes the other one.

## Problem 2

Two students conduct a study to investigate the relationship between forearm length and height:

- Maria measures the subjects in centimeters. In a scatterplot of the data she sees a linear relationship between the variables, so she calculates the correlation coefficient. She determines that r = 0.86.
- John measures the same subjects in inches. He also calculates the correlation coefficient.

What do you expect the correlation will be for John's measurements?

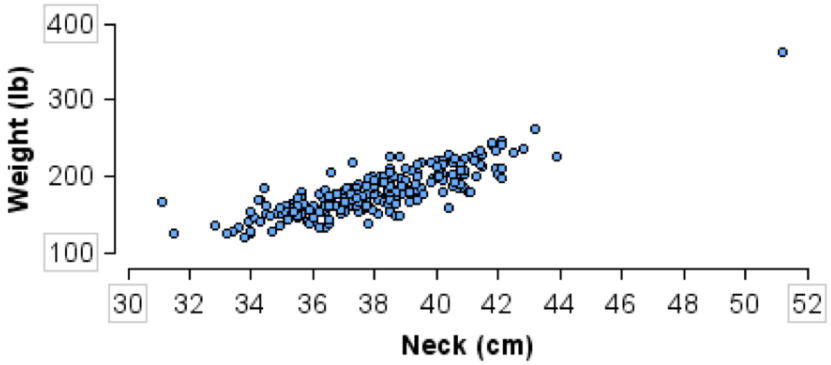
a. John's correlation will be approximately 0.86(0.61) = 0.52 because one centimeter = 0.61 inches.

b. John's correlation will be approximately r = 1 because there is a very strong relationship between John's measurements and Maria's measurements.

c. John's correlation will be approximately 0.86 because the pattern in the data will be the same.

**Problem 3**

This scatterplot shows the relationship between neck measurement (cm) and weight (lb) for 252 men. The correlation is 0.83.




If we change the weights from pounds into kilograms, what is the effect on the correlation?

a. The correlation is close to 1 because pounds and kilograms are strongly associated.

b. The correlation is approximately 0.83/2.2 = 0.04 because 1 kg is 2.2 lbs.

c. The correlation is approximately 0.83 because the pattern in the data will not change.

**Problem 4**

True or False, and explain: if the correlation coefficient is 0.90, then 90% of the points are highly correlated.

**Problem 5**

In a study of 6-to 11-year-old elementary school children, researchers find a strong positive association between reading level and weight. They found out that heavier children tend to have higher reading levels. To increase reading scores, should the school develop a strategy to help children gain weight? Why or why not?

**Problem 6**

For a certain data set, $r = 0.57$. Say whether each of the following statements is True or False, and explain briefly. If you need more information, say what you need, and why.

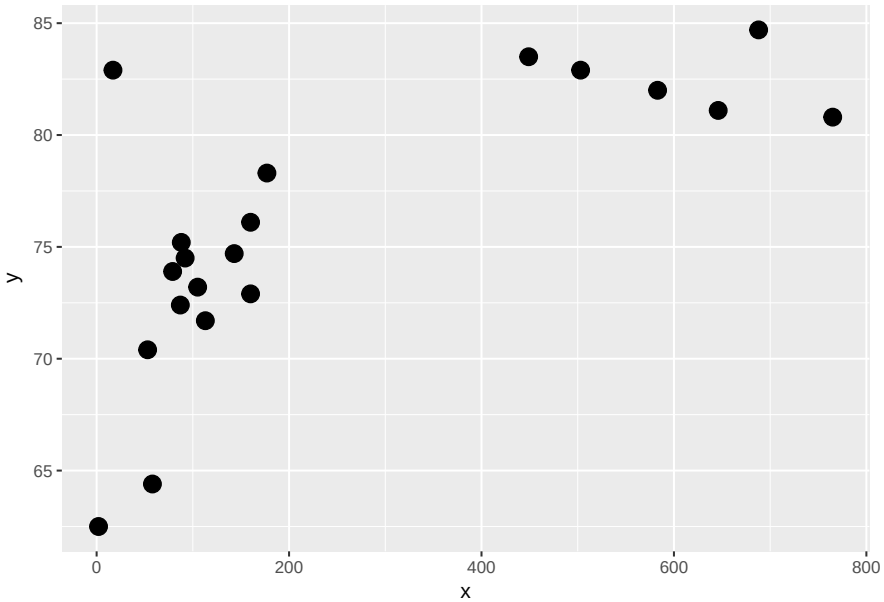
a. There are no outliers.

b. There is non-linear association.

**Problem 7**

For the 20 countries with the largest population for 2009 the scatterplot shows the following form.

- $x$ = internet users per 1000 people
- $y$ = life expectancy (years) (World Almanac Book of Facts, 2009)




a. The association between these two variables is positive. Explain what this means for this context.

b. The correlation coefficient is 0.72, which is strong. Larger numbers of internet users per 1,000 correlate with longer life expectancy. Someone who confuses correlation with

causation might suggest that an easy way to improve a country's life expectancy is to get more people onto the internet, which is a ridiculous cause-and-effect statement. Identify a lurking variable that might be explaining the strong association between life expectancy and the number of internet users per 1,000.

**Problem 8**

For a sample of cities in the U.S. the correlation between the murder rate and number of police officers employed by the city is r = 0.72. What can we conclude?

a. There is a strong positive linear relationship between murder rate and size of the police force.

b. Cities are responding to an increase in the murder rate by hiring more police officers.

c. The size of the city is a possible lurking variable explaining the association between murder rate and size of the police force.

d. Both (a) and (c) are true.

e. None of the above are true.

**Problem 9**

In an introductory course to Statistics, small discussion sections are led by teaching assistants. As part of a study, at the second-to-last lecture one term, the students were asked to fill out anonymous questionnaires rating the effectiveness of their teaching assistants (by name), and the course, on the scale:

`1 = poor,  2 = fair,  3 = good,  4 = very good,  5 = excellent`

The following statistics were computed:

- The average rating of the assistant by the students in each section.
- The average rating of the course by the students in each section.
- The average score on the final for the students in each section.

Results are shown below (sections are identified by letter).

a. Use R to create three vectors: `assistant`, `course`, and `final`

b. Find the correlations: 1) assistant rating -vs- course rating, 2) assistant rating -vs- final score, and 3) course rating -vs- final score.

c. Plot scatter diagrams of: 1) assistant rating -vs- course rating, 2) assistant rating -vs- final score, and 3) course rating -vs- final score.

| section | assistant | course | final |
|---|---|---|---|
| A | 3.30 | 3.50 | 70.00 |
| B | 2.90 | 3.20 | 64.00 |
| C | 4.10 | 3.10 | 47.00 |
| D | 3.30 | 3.30 | 63.00 |
| E | 2.70 | 2.80 | 69.00 |
| F | 3.40 | 3.50 | 69.00 |
| G | 2.80 | 3.60 | 69.00 |
| H | 2.10 | 2.80 | 63.00 |
| I | 3.70 | 2.80 | 53.00 |
| J | 3.20 | 3.30 | 65.00 |
| K | 2.40 | 3.30 | 64.00 |

**Problem 10**

*(Refer to the previous question).* The data are section averages. Since the questionnaires were anonymous, it was not possible to link up student rating with scores on an individual basis. Student ability may be a confounding factor. However, controlling for pre-test results turned out to make no difference in the analysis. Each assistant taught one section. True or False, and explain:

a. On the average, those sections that liked their TA more did better on the final.

b. There was almost no relationship between the section's average rating of the assistant and the section's average rating of the course.

c. There was almost no relationship between the section's average rating of the course and the section's average score on the final.

**Problem 11**

A teaching assistant gives a quiz with 10 questions and no partial credit. After grading the papers, the TA writes down for each student the number of questions the student got right and the number wrong. The average number of right answers is 6.4 with and SD of 2.0; the average number of wrong answers is 3.6 with the same SD of 2.0. The correlation between the number of right answers and the number of wrong is: *(Explain your answer)*

a. 0

b. -0.50

c. 0.50

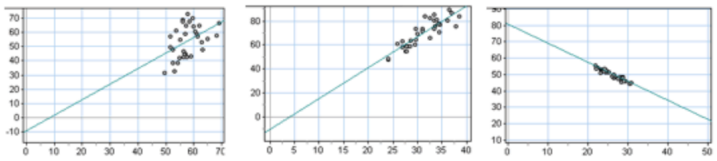
d. -1

e. 1

f. can't tell without the data

## Problem 12

Consider the following graphs



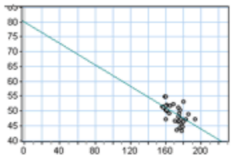
Graph A



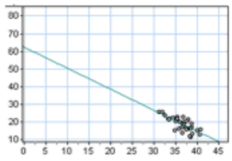
Graph B



Graph C



Graph D



Graph E

a) Below are the correlation coefficients and regression equations for the 5 scatterplots pictured above. Match the plots with their correlation coefficients.

i. -0.95

ii. -0.73

iii. -0.54

iv. 0.45

v. 0.88

b) Type the letter of the scatterplot that matches each regression equation.

i. $Y = -10.5 + 1.1X$

ii. $Y = -10.5 + 2.6X$

iii. $Y = 62 + (-1.2)X$

iv. $Y = 80 + (-1.2)X$

v. $Y = 80 + (-0.2)X$

## Problem 13

Consider the following data on `age` and `price` for 11 Orions (a type of car). Price is measured in hundreds of dollars. In this case, `age` is the predictor variable, and `price` is the response variable.

```
age <- c(5, 4, 6, 5, 5, 5, 6, 6, 2, 7, 7)
price <- c(85, 103, 70, 82, 89, 98, 66, 95, 169, 70, 48)
```

a. Use R to calculate the means, SDs, and correlation. Display these values.

b. Calculate the slope and y-intercept of the regression line. And write its equation.

c. Graph the regression equation and the data points. You can use `plot()` to create the scatterplot, and then `abline()` to add the regression line (by providing the intercept and the slope)

d. Describe the aparent relationship between age and price of Orions.

e. How would you interpret the slope of the regression line?

f. Use the regression equation to predict the price of a 3-year-old Orion and a 4-year-old Orion.