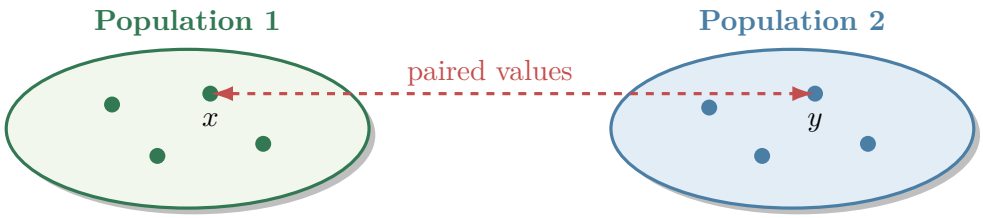
MAT220 Lecture 18: Linear Correlation

Contents

In the previous two lectures, we performed hypothesis tests on two populations simultaneously. In this lecture, we will use paired data in a completely different way: instead of testing population parameters against each other, we look at descriptive statistics of the paired data. We will concern ourselves with *linear correlation*, which asks whether or not two data sets display a pattern resembling a line. We will start our discussion intuitively (that is, visually) and then we will look at ways to describe linear relationships mathematically.

1 What is Correlation?

Recall that in Lecture 16 we discussed the notion of *paired* populations. In short, these are populations in which each value in one population has a natural counterpart or partner in the other:



Paired data can arise in several ways:

- **successive experiments on the same group,** such as reaction times before and after caffeine;
- **different experiments on the same object,** such as heart-rate readings from a chest strap and an Apple Watch during the same run;
- **naturally mirrored or matched objects,** such as first-born and second-born twins, or the left and right feet of the same person.

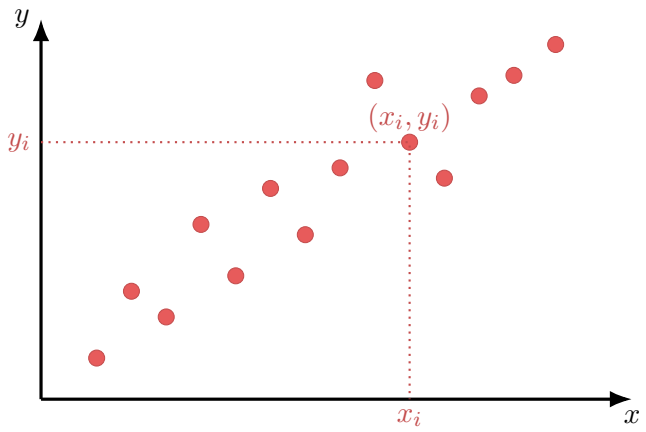
Previously, we used the pairing to form differences of the random variables, which we wrote $D = X - Y$. We then used this to test whether or not X and Y have different means. Today, we are simply going to study *descriptions* of paired data. We start with a visual description known as a scatter diagram.

1.1 Scatter diagrams

Suppose that we take a sample of n pairs of data. This can be written as a list:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

A *scatter diagram* is a visual representation of this data in which every pair (x_i, y_i) is plotted as an individual point in two-dimensional space. The horizontal axis represents the x-variable and the vertical axis represents the y-variable, so that each point represents a data pair:



In reality, data of this type will occur whenever we measure the same object in two different ways. The hope is that by plotting the paired observations, we can visually inspect the data to see if there are any patterns. As a mathematical setup, we often try to explain the variable Y in terms of the variable X. For that reason, we call x the *explanatory variable*, and we call y the *response variable*.[1]

Scatter diagrams

A **scatter diagram** is a graph of paired quantitative data in which each pair (x_i, y_i) is plotted as one point.
The variable x is called the **explanatory variable**, while y is called the **response variable**.

The terminology suggests the type of question we are asking. We suspect that changes in x may help to explain why y changes. For example, suppose x is the speed of a car and y is its stopping distance. It is reasonable to suspect that the speed helps explain the stopping distance, and that the stopping distance responds to changes in speed according to some underlying rule.

1.2 Linear correlation

The word "correlation" means that two variables display some kind of pattern together. In principle, the pattern between two variables could have many different shapes, i.e. two sets of data may be correlated in many different ways. In this lecture, we will study the simplest possibility: *linear* correlation. Roughly speaking, the data has linear correlation when the pattern of points approximately follows a straight line.

Linear correlation

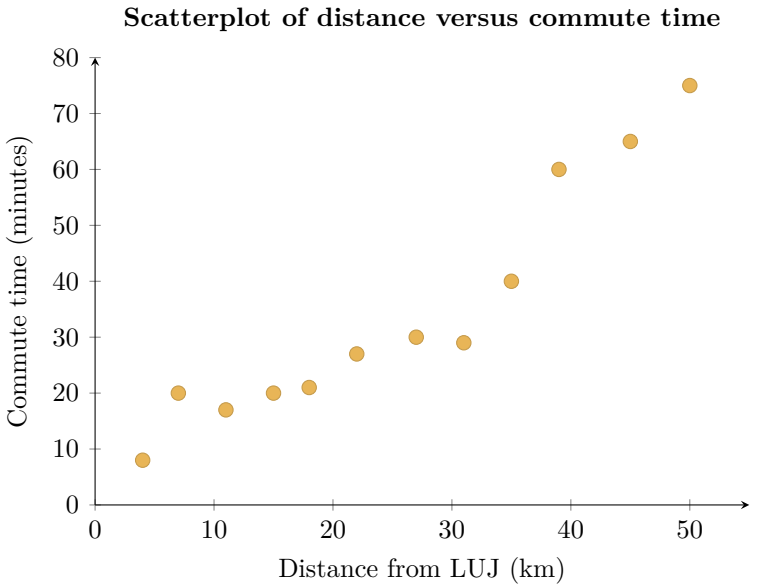
Paired data displays a **linear correlation** when the scatter diagram displays an overall pattern that looks like a straight line.

Linear relationships are extremely common. Examples include the population and GDP of geographical regions, the force required to extend a spring, and oxygen consumption compared with energy expenditure. Later, in Lecture 19, we will introduce *linear regression*: a method for finding the line that best approximates a linear pattern.

As an example, suppose that we took a survey of 12 students, and recorded the distance from their home to campus, and also the time it takes them to commute here. Of course, this data is naturally paired: both values are connected due to the fact that they relate to the same people. The data and the corresponding scatter diagram are depicted below.

[1] There are other common names for these two variables in the literature. Sometimes, x is known as the *independent* variable and y is known as the *dependent* variable.

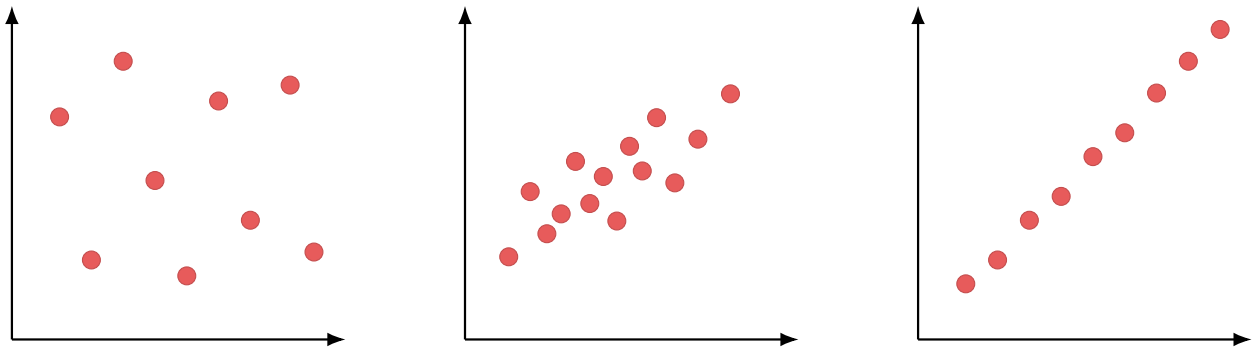
Student	Distance (km)	Time (min)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75



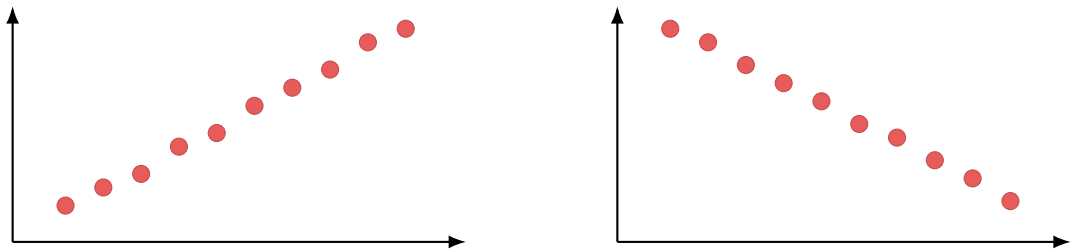
As you can see, the data points display a clear linear pattern: those students who live close to campus have a short travel time, and those students who live far away have a longer travel time. Of course, this data does not arrange itself as a perfect straight line, and this is to be expected: there are many small factors that influence travel time, such as train lines, travel methods and so on. However, based on the data, we can still see a general pattern.

1.3 The strength and sign of a correlation

Generally speaking, some scatter diagrams may contain no obvious patterns at all, whereas others very much resemble a line. In natural language, we describe these differences using the "strength" of the correlation. Roughly speaking, a correlation is called "strong" if the pattern is quite clear, and a correlation is called "weak" if there is still *some* pattern, but not very much:



In addition to this, there is another important feature of linear correlation worth mentioning. Consider the two images below:



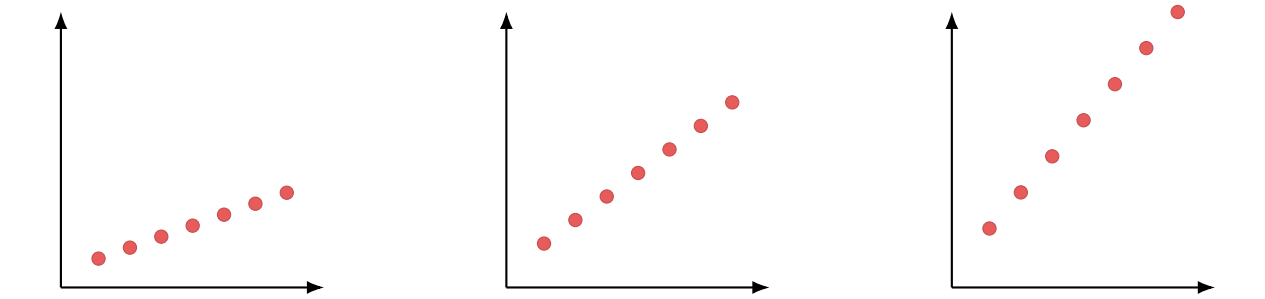
Both of these display a strong linear correlation, yet they are clearly different. In the graph on the left, x and y increase together: when x is low so is y, and when x is high so is y. In contrast, in the graph on the right, the relationship is flipped: when x is low y is high, and when x is high, y is low. In other words, the lines that fit these data sets have gradients of different *signs*. This observation motivates the following terminology.

The sign of a correlation

The *sign* of a correlation describes the direction of the general trend:

- **positive correlation:** x and y tend to increase together, so the general trend has a positive gradient;
- **negative correlation:** as x increases, y tends to decrease, so the general trend has a negative gradient.

Note that here, the words "positive" and "negative" only refer to the *sign* of the pattern, not to the gradient itself. For instance, all three graphs below display a positive linear correlation, despite having different gradients:



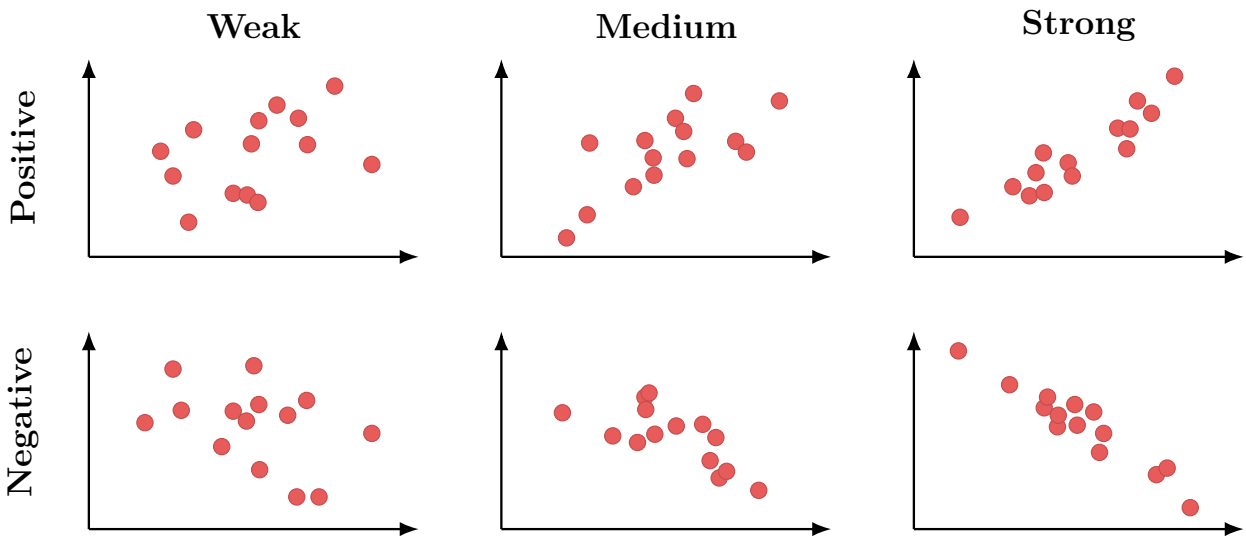
Perfect positive correlation

Perfect positive correlation

Perfect positive correlation

1.4 Sign versus strength

Sign and strength are simply two different features of linear correlation. As the picture below summarizes, we can find weak, medium, or strong patterns of either sign.



2 The Correlation Coefficient

2.1 The sample correlation coefficient

Although intuitive, natural language terms like "weak" and "strong" are somewhat vague. To resolve this, we will now introduce a number that quantifies how strong a correlation is. Simply put, we are now going to define a number that measures how strong a correlation is: the stronger the correlation, the higher the absolute value of this number, and the weaker the correlation, the lower the absolute value. This number is typically known as the *sample correlation coefficient*, and is written r. By design, it measures both the strength and the sign of a linear relationship between the variables x and y.

The sample correlation coefficient

The **sample correlation coefficient** r is a unitless number satisfying

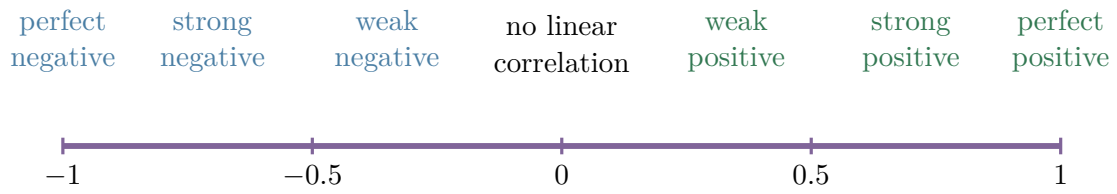
$$-1 \leq r \leq 1.$$

Its sign records whether the linear correlation is positive or negative, while the absolute size of r measures the strength of the linear relationship.

The term "sample correlation coefficient" can be broken down into its three pieces:

- the term "coefficient" tells us that r is a number,
- the term "correlation" refers to the pattern between x and y, and
- the term "sample" refers to the fact that the number r is built from our sample data (x_i, y_i).

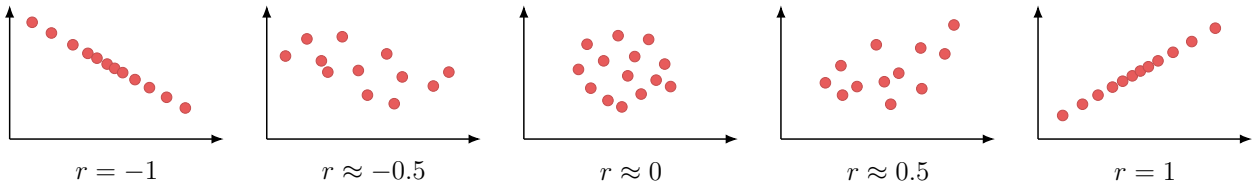
Therefore, the term "sample correlation coefficient" really just means "a number that measures how much our sample follows the pattern of a line". Using the number r, we can be more precise about the strength of a linear correlation:



In words, the important reference values are:

- $r = 1$ means that the data is arranged perfectly on a straight line with a positive gradient,
- $0 < r < 1$ describes some positive linear correlation,
- $r = 0$ says that the data has no linear correlation at all,
- $-1 < r < 0$ describes some negative linear correlation in the data, and
- $r = -1$ indicates perfect negative linear correlation.

The following diagrams illustrate data with several different approximate values of r:



Even if there is a precise mathematical rule relating x and y, real-world data will rarely follow that rule perfectly. Measurements contain noise, individual differences, and small errors. Perfect linear correlation is therefore best viewed as an idealized theoretical limit.

Although we will not discuss it directly in this course, it is worth mentioning that there is also a population parameter that measures linear correlation. As with the mean or the standard deviation, we denote the population parameter by a Greek letter, this time we use the letter ρ:

	Sample	**Population**
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Correlation coefficient	r	ρ

The sample correlation coefficient r measures the linear relationship between the pairs of points in a given sample. In contrast, the population correlation coefficient ρ describes the true correlation between all pairs in the full population.

2.2 The formula for r

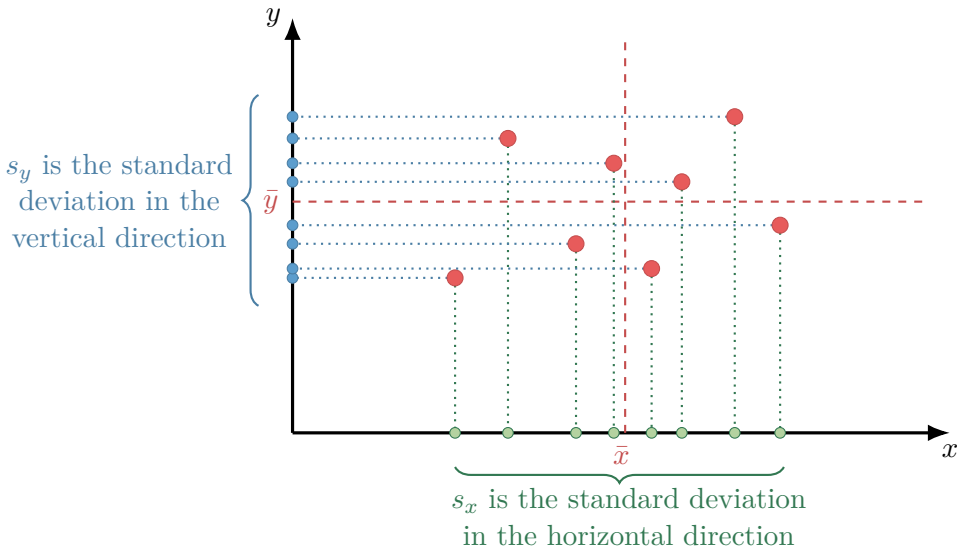
The sample correlation coefficient is defined using the formula:

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

At first glance, this formula looks overwhelmingly complicated. To better explain what is going on here, we will break down the formula step by step. Notice first that the denominator of r has two sample standard deviations:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{and} \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

As a reminder: here we are working with *paired* data, meaning that every data point in our sample is of the form (x_i, y_i). Each pair can be separated into its two coordinates and considered individually:



So, in the formula for r we are somehow dividing by the individual spreads of x and y. In the numerator of r, we have a new term, known as the *covariance*:

$$\text{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

In fact, this is the sample version of the covariance we mentioned in Lecture 16 when discussing the variance of paired samples. Back then, we merely commented on its existence as a measure of the interrelatedness of x and y. We will now talk at length about what this really means.

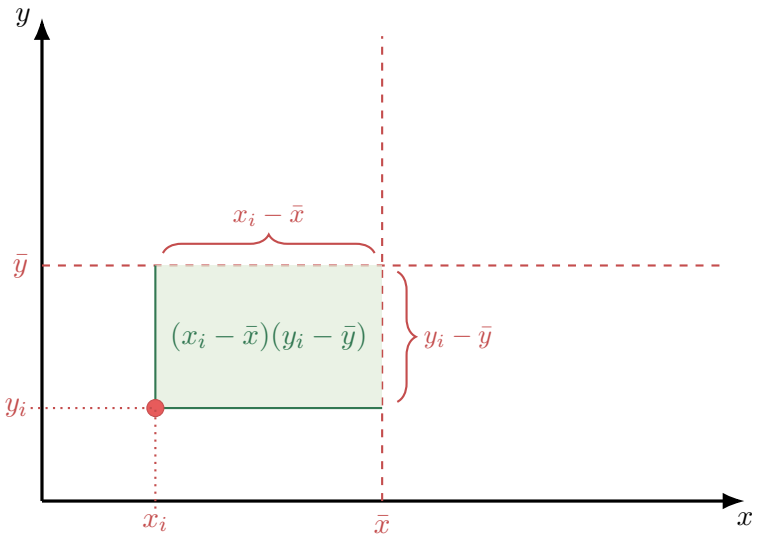
2.3 Understanding covariance

Let's examine the formula for covariance:

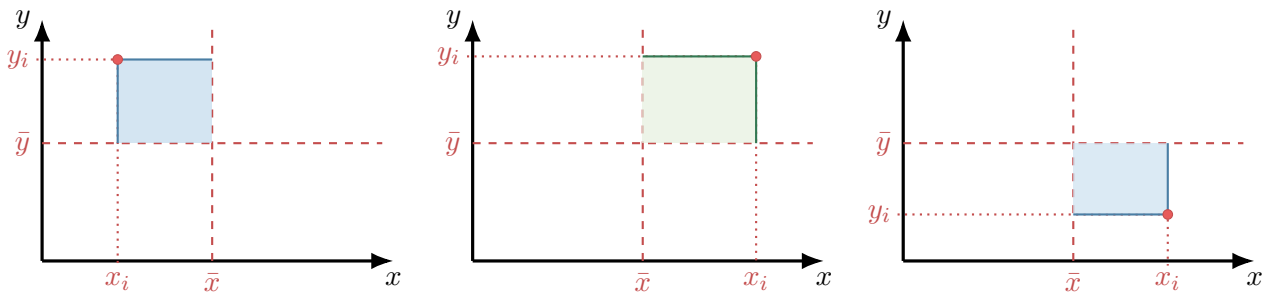
$$\text{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

What's going on here? Firstly, we may notice that on the very outside we are performing a sum over many quantities, and then we divide by $n - 1$. This part of the formula looks similar to that of sample variance: in fact, it resembles a sort-of average. So, what are these quantities, exactly?

The individual parts $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are actually measuring distances between points: they measure how far x_i is from $\bar{x}$ and how far y_i is from $\bar{y}$. These are two lengths, and multiplying them together forms an *area*:



In the picture above, we have deliberately drawn our data point (x_i, y_i) in the bottom left quadrant of the chart. In this situation, x_i is to the *left* of $\bar{x}$, meaning that $x_i < \bar{x}$ and therefore $x_i - \bar{x} < 0$. Similarly, the component y_i is *below* the mean $\bar{y}$, meaning that $y_i < \bar{y}$ and therefore $y_i - \bar{y} < 0$. Therefore our expressions $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are distances that carry a *sign*.[2] If both of these signed distances are negative, then multiplying them together gives an overall positive answer. As for the other three quadrants, we have different overall signs of the areas:



The sign of each rectangle depends only on which side of the two means the point lies:

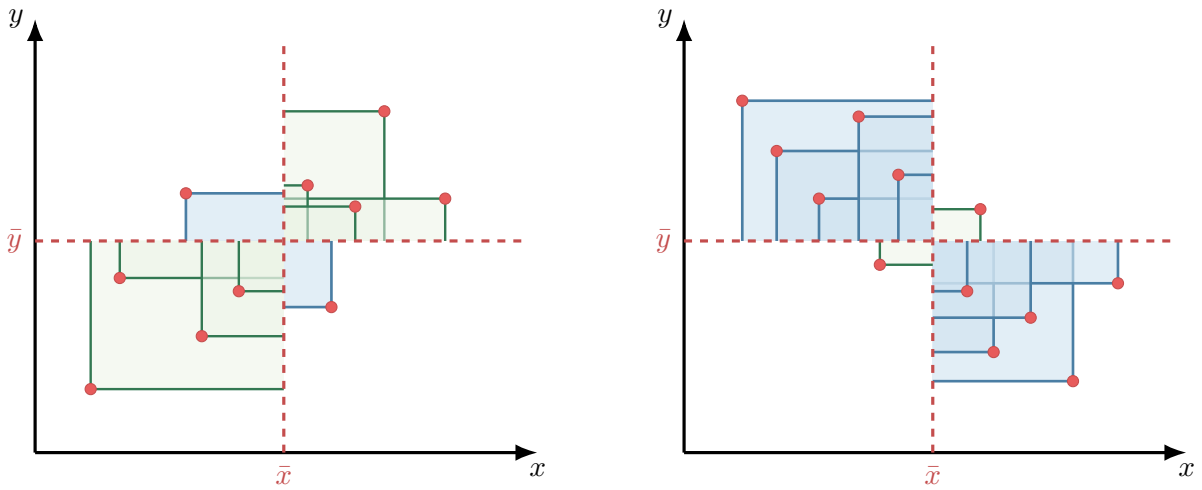
Location of (x_i, y_i)	$(x_i - \bar{x})(y_i - \bar{y})$	Signed Area
$x_i < \bar{x},\ y_i < \bar{y}$	$(-)(-)$	positive
$x_i < \bar{x},\ y_i > \bar{y}$	$(-)(+)$	negative
$x_i > \bar{x},\ y_i > \bar{y}$	$(+)(+)$	positive
$x_i > \bar{x},\ y_i < \bar{y}$	$(+)(-)$	negative

If $x_i = \bar{x}$ or $y_i = \bar{y}$, then the signed area is 0.

[2]These are often called *signed deviations* from the mean.

The covariance adds together the signed areas of all of these rectangles, and then divides by $n-1$. In other words, covariance is an average measure of whether the two variables lie on the same side of their means, and how far they lie from those means.

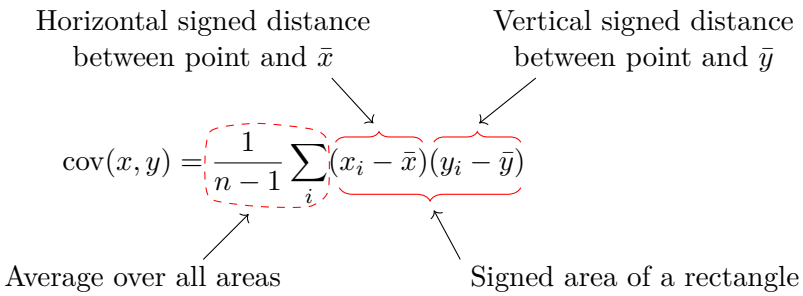
In fact, this "sort-of average of signed area" allows us to recover the sign of a possible correlation. In the diagram below, we have created the signed rectangular areas for each of the data points. Here we represent the areas with a positive sign using green, and the areas with a negative sign using blue.



Notice that the diagram on the left is somewhat positively correlated: most of the data lies in the bottom left and top right quadrants of the diagram. Visually, there is more green in the diagram than there is blue, and therefore the covariance should be positive. In contrast, the diagram on the right is negatively correlated: most of the data points are in the top left or bottom right of the diagram. Visually, the signed rectangles are mostly blue, meaning that the covariance will be negative. To summarize:

- if the diagram contains more green area than blue, then the sum of the signed areas is positive, and $\text{cov}(x, y) > 0$,
- if the diagram contains more blue area than green, then the sum of the signed areas is negative and $\text{cov}(x, y) < 0$, and
- if the positive and negative areas largely cancel in the sum, then the covariance is close to 0.

Going back to the formula, each symbol can be broken down as follows:

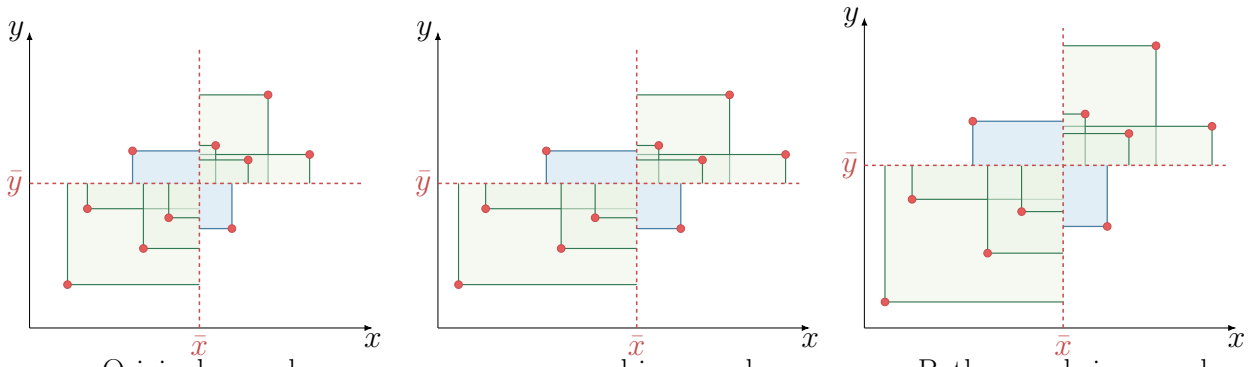


What covariance measures

Covariance measures whether the two variables tend to lie on the same side of their respective means.

2.4 Understanding the formula for r

By construction, the formula for covariance cleverly uses areas of rectangles to keep track of the sign of any (potential) linear correlation. However, covariance on its own still depends on the overall spread of the data in question. To see this, let's take a data set and manually increase the spread in both the x and y directions:



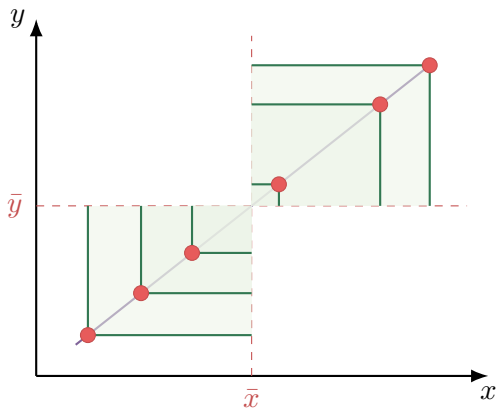
Notice that as we increase the spread of the data, the average distance of each point from either $\bar{x}$ or $\bar{y}$ will also increase, meaning that the average area of each rectangle will also increase. However, generally speaking, increasing the spread in this way doesn't turn a green rectangle into a blue one, or vice versa. If we wanted to care only about the interrelatedness of green and blue rectangles, rather than their raw size, then we could divide by the spread of the data to get rid of their absolute sizes. This is the formula for r:

$$r = \frac{\text{cov}(x, y)}{s_x s_y}.$$

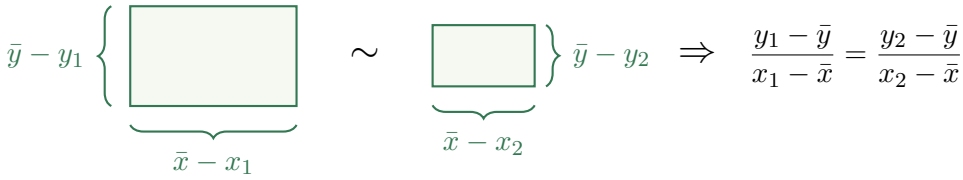
Roughly speaking, the standard deviations s_x and s_y measure the spread of the data in the horizontal and vertical directions, respectively. If we divide the covariance by these two terms, then we get rid of the overall scale of the diagram above, so that increasing the spread of the data doesn't actually change the answer. Instead, all that matters is the relative pattern of the areas formed from the data.

2.5 Recovering $r = 1$ for Perfect Positive Correlation

Suppose now that we have data that is perfectly arranged on a line. We will now demonstrate how our formula for r gives a value of $+1$. In the image below, we have plotted six points on a straight line with positive gradient, and have drawn all of their signed areas to the side.



Geometrically, all of these rectangles have the same *aspect ratio*:



We will call this aspect ratio k. If all the points (x_i, y_i) lie on the same line, then each of these ratios equals k.[3] Rearranging, for all $i = 1, \ldots, n$ we have:

$$\frac{y_i - \bar{y}}{x_i - \bar{x}} = k \quad \Rightarrow \quad y_i - \bar{y} = k(x_i - \bar{x}).$$

If we substitute this into the formula for covariance, we happen to recover an expression containing the sample variance for x:

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) k (x_i - \bar{x}) = \frac{k}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = k s_x^2.$$

Similarly, replacing all occurrences of $y_i - \bar{y}$ with $k(x_i - \bar{x})$ in the formula for s_y happens to give:

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} k^2 (x_i - \bar{x})^2} = k s_x,$$

where the last equality uses $k > 0$. Substituting into the formula for r gives

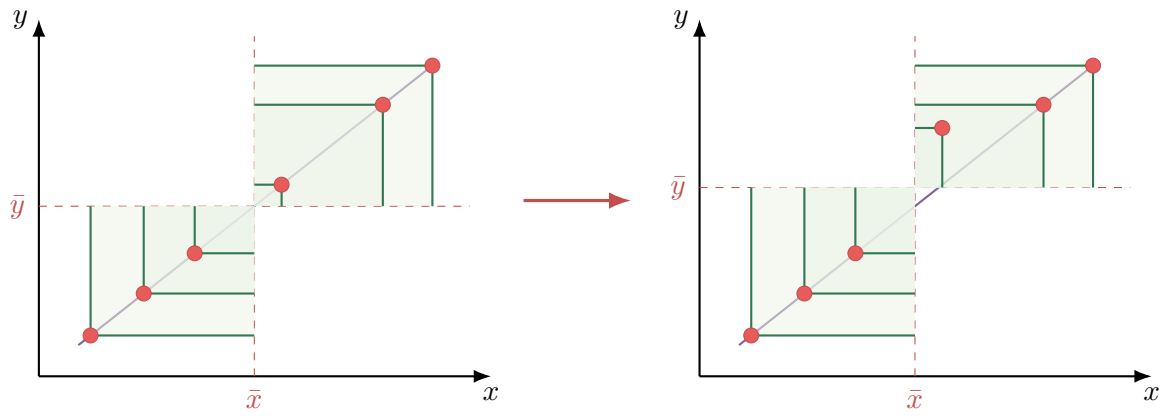
$$r = \frac{k s_x^2}{s_x (k s_x)} = 1,$$

as expected.

[3] Technically, this does not work whenever the point has either $x_i = \bar{x}$ or $y_i = \bar{y}$. In those cases, the "areas" would collapse to a one-dimensional line.

2.5.1 What happens when one point moves?

To get a better idea of what r is doing geometrically, let's take a point and shift it up a bit, so that the data is no longer linear:



We would expect that doing this would decrease r from 1 to something slightly smaller. Visually, we can see that manually moving one point off the line vertically will drag up the overall mean $\bar{y}$. Therefore, all of the distances $y_i - \bar{y}$ will also change. In the process, these rectangles will no longer have the same aspect ratio: in the image above, the rectangles now have different shapes.

Algebraically, shifting a single point vertically adds a small piece to the y component of some fixed point, e.g. $(x_4, y_4) \to (x_4, y_4 + \varepsilon)$, where ε is some positive number. Roughly speaking, making this change to one point will make the covariance grow a little bit, but the formula for the standard deviation s_y contains lengths *squared*, i.e. pieces of the form $(y_i - \bar{y})^2$. When we shift a single point by ϵ, this overall small change cascades into a relatively big change because of the exponent. This causes the s_y term to grow faster than the $cov(x, y)$ term, and therefore the ratio r shrinks to be smaller than 1.

2.6 Example: calculating r

We now return to the commute-time data from Section 1. The sample means are

$$\bar{x} \approx 25.333 \qquad \text{and} \qquad \bar{y} \approx 34.333.$$

In order to compute r, we need to sum over several types of terms. The table below records the relevant information:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
4	8	-21.333	-26.333	455.111	693.444	561.778
7	20	-18.333	-14.333	336.111	205.444	262.778
11	17	-14.333	-17.333	205.444	300.444	248.444
15	20	-10.333	-14.333	106.778	205.444	148.111
18	21	-7.333	-13.333	53.778	177.778	97.778
22	27	-3.333	-7.333	11.111	53.778	24.444
27	30	1.667	-4.333	2.778	18.778	-7.222
31	29	5.667	-5.333	32.111	28.444	-30.222
35	40	9.667	5.667	93.444	32.111	54.778
39	60	13.667	25.667	186.778	658.778	350.778
45	65	19.667	30.667	386.778	940.444	603.111
50	75	24.667	40.667	608.444	1653.778	1003.111
Total		0	0	**2478.667**	**4968.667**	**3317.667**

Using the simplified computational formula,

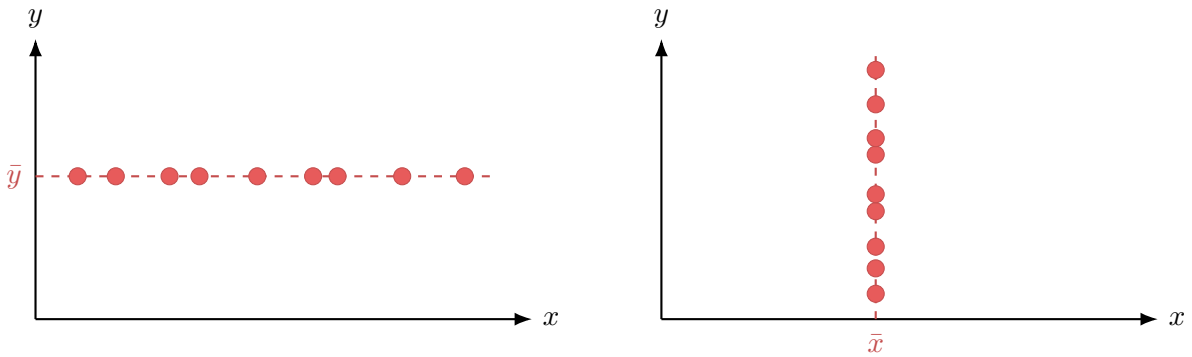
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{3317.667}{\sqrt{2478.667}\sqrt{4968.667}} \approx 0.945.$$

This is a strong positive linear correlation, meaning that students who live farther from LUJ tend to have longer commute times.

3 Limitations of Correlation

3.1 Cases where r is undefined

Consider the following two data sets:



Clearly, these data sets both lie on a straight line. However, it turns out that the sample correlation coefficient r is undefined. In either case, one of the two directions has *zero* spread, meaning that

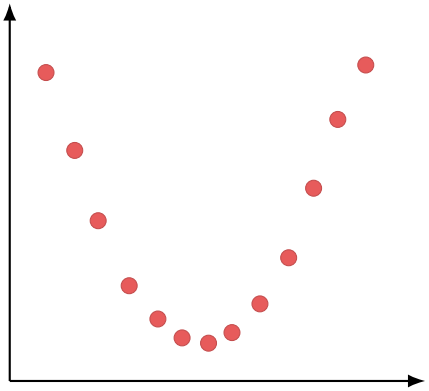
either every y_i equals $\bar{y}$ or every x_i equals $\bar{x}$. This causes the covariance $\text{cov}(x, y)$ and either s_y or s_x to vanish, giving:

$$r = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{0}{0},$$

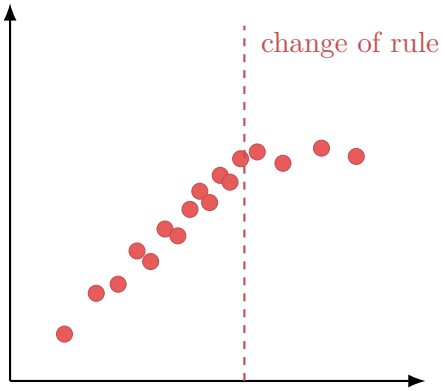
which is undefined.

3.2 Non-linear correlation

The correlation coefficient r only measures *linear* correlation. In particular, some data sets may be correlated in other ways, yet still have a small value of r. In other cases, it may be that the relationship is linear for a while, yet then changes to some other type of relationship:



Curved relationship



Phase transition

In scientific applications, a relationship may be approximately linear only up to a certain point, after which a new rule takes over. In situations like this, the correlation coefficient may return a high value anyway, even though it is not technically representative of the full system.

3.3 Correlation without causation

Just because there is an apparent correlation in data does not mean that there is necessarily a causal link between the two variables. For instance, if X and Y are correlated, then

- X may cause Y, or
- Y may cause X, or
- X and Y may be correlated by pure coincidence, or
- there may be a third variable that influences both X and Y.

The slogan here is:

Correlation does not imply causation.

As an example, ice cream sales are often correlated with suncream sales. This does not mean that ice cream sales cause suncream sales or vice versa; instead, the weather is a third variable that influences both.

Correlation does not imply causation

The value of r can describe the existence, sign, and strength of a linear pattern. However, it cannot give that pattern a causal interpretation.