Linear regression

numeric variables

A numeric variable is an "attribute or value" that is described using a number. That definition probably doesn't help much, but some examples should.

Think about the cost of a pound of ground beef, or the number of grams of jello pudding contained in a box. Both of these can be described by a number, and they both might vary from day-to-day or box-to-box.

Take your pulse for a minute and count the number of heartbeats. Set a timer on your phone to count the number of seconds it takes to ride your bike one mile.

There are many situations where two numerical variables are correlated. If you ride the same route every day on your bicycle, riding faster will make the time decrease, and your heart rate will almost certainly go up. There is a negative correlation between beats per minute and elapsed time.

Years are a very common variable. These might be adjusted, say, only counting years since the year 2000, or they might be the full C.E. value, like 2022. It turns out not to matter because of *subtracting the mean*, as we will see.

bivariate data

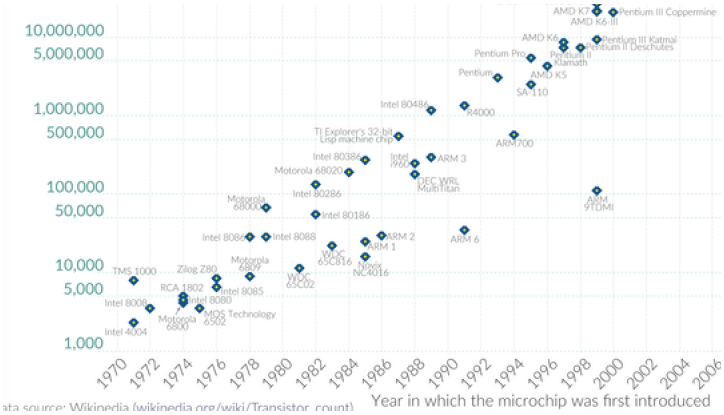
A well-known textbook of Statistics (Bock *et al*) shows an example of rollercoasters on its cover. It is perhaps not too surprising that the maximum height of a roller coaster is correlated with the top speed produced during the run.

Part of a college admissions officer's job is to admit students who are more likely to do well in the future, perhaps because they have good grades in high school, e.g. a GPA closer to 4 than to 2.

She might ask whether there is a positive correlation between grades in high school and in college. Another comparison might use test scores in admission exams rather than high school GPA. Which does a better job of predicting "success"?

Gordon Moore, a famous computer pioneer, noted that the number of transistors carried on an integrated circuit (IC) in computers doubles about every two years. This was called Moore's Law.

Since the IC chips stay roughly the same size, for this trend to continue, the transistors on a chip must be getting smaller. The current size is around 5 nanometers, only about 25 times the size of the silicon atoms (0.2 nm) from which circuits are built. Obviously, this trend will not continue much longer, but it was true for at least 50 years.



https://en.wikipedia.org/wiki/Moore%27s_law

We note in passing that that the values plotted above on the y-axis have been transformed by taking the logarithm of each data point. This is called a *semi-log* plot. This method is commonly used when the rate of change in the y variable is exponential with time, which means that it doubles every n years.

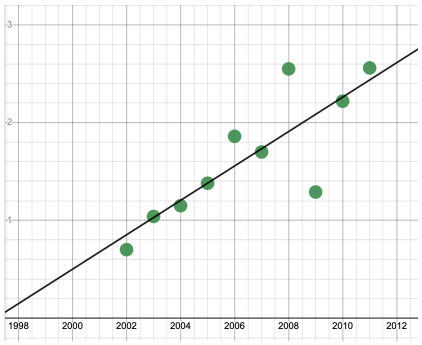
A third example is the price of a commodity, like gasoline.

AVERAGE GAS PRICES

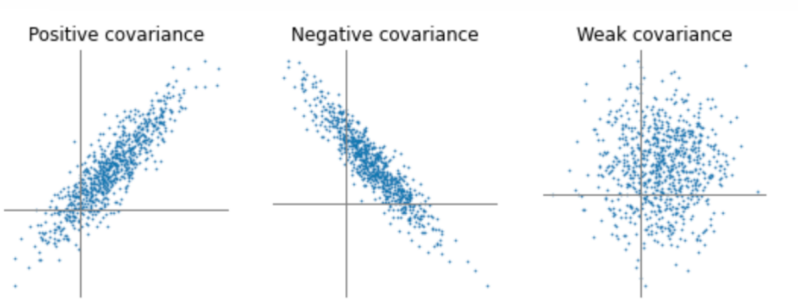
Columbia, South Carolina

Year	Gas Price	Year	Gas Price
2002	$0.70	2007	$1.70
2003	$1.04	2008	$2.55
2004	$1.15	2009	$1.29
2005	$1.38	2010	$2.22
2006	$1.86	2011	$2.56

If we plot each year on the x-axis and the price on the y-axis, it looks like this:



For plots like this there are three general possibilities: positive correlation, negative correlation, and uncorrelated.



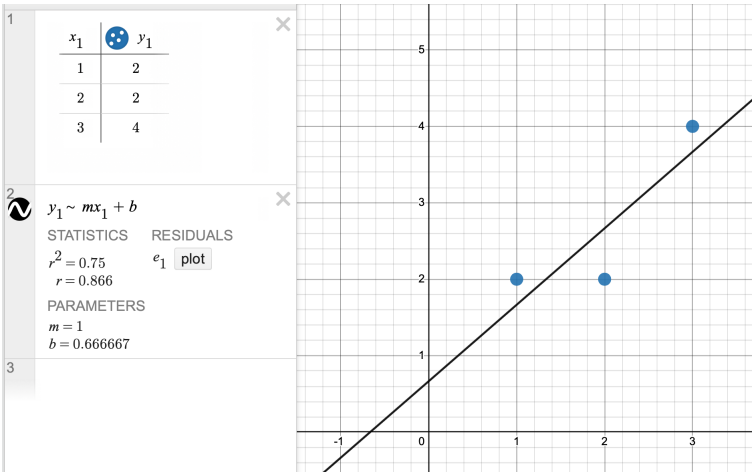
A technical word for this is *covariance*. Covariance is quantitative, it is described by a statistic (a number) which indicates whether the correlation is excellent or very good, moderate or just plain terrible.

toy example

Let us make a very simple example, in order to understand how correlation is measured. Start with three pairs of (x, y) values: (1,2), (2,2), (3,4).

We want to find an equation so that we can predict the value of y if we know x. It won't be perfect, but that's OK.

Open Desmos and plug them into a table, then make a linear model. Do this by typing $y_1 \sim mx_1 + b$. That's a subscript 1 on both the x and y, entered by first typing the underscore _ and then the 1.



Desmos says that the linear regression model gives a line with a slope of 1 and a y-intercept of 2/3. We want to know how the calculation was done. It's pretty easy.

If you look at the points and try to guess where the center of the mass of points lies, you would probably say it is near $x = 2$ and maybe $y =$ something a bit larger than 2.

mean

The first big idea in linear regression is to center the data, to move this average point to the origin. This is done be subtracting the average or *mean* values for both x and y.

We need the two means, for the x and y values. One symbol for the mean is to put a little bar over the letter, and in talking about them say "x-bar" and "y-bar".

$$\bar{x} = \frac{1+2+3}{3} = 2 \qquad \bar{y} = \frac{2+2+4}{3} = 8/3$$

Sometimes, we use another convention, which is to call the mean of x, $\mu(x)$.

subtracting the mean to center the data

Take each x value in the list of values and subtract the mean. Do not add them together yet but just keep them in order in a list.

$$x_1 - \bar{x} = 1 - 2 = -1$$
$$x_2 - \bar{x} = 2 - 2 = 0$$
$$x_3 - \bar{x} = 3 - 2 = 1$$

You may notice that if we *did* add them together, the result would be zero. Now do the same for y.

$$y_1 - \bar{y} = 2 - 8/3 = -2/3$$
$$y_2 - \bar{y} = 2 - 8/3 = -2/3$$
$$y_3 - \bar{y} = 4 - 8/3 = 4/3$$

It may be worthwhile to plot the points generated by this procedure. If we subtract $\bar{y}$ from each value in a list of y-values, what is the mean of the adjusted values?

What happens to a bunch of points (x, y) if we subtract $\bar{x}$ from their x-values and $\bar{y}$ from their y-values?

multiply, and add

The second big idea is a technique of linear regression called the dot product. Take the corresponding values from each list and multiply, then add the results together.

For example, the first term is $(x_1 - \bar{x}) \cdot (y_1 - \bar{y})$. I'm using $\cdot$ as the symbol for multiplication.

$$(-1 \cdot -2/3) + (0 \cdot -2/3) + (1 \cdot 4/3)$$
$$= 2/3 + 0 + 4/3 = 6/3 = 2$$

Th result of this calculation is often called SS_{xy}. (Not because it's *so sexy*). SS is actually short for "sum of squares".

We're almost there! Go back to the first list and multiply each term by itself. The first one is $(x_1 - \bar{x})^2$. Add them up.

$$-1 \cdot -1 + 0 \cdot 0 + 1 \cdot 1$$

$$= 1 + 0 + 1 = 2$$

This is often called SS_{xx}.

equation

Divide the first result by the second (SS_{xy}/SS_{xx}). This is the slope of the line that gives the best fit by linear regression.

$$m = \frac{SS_{xy}}{SS_{xx}} = \frac{2}{2} = 1$$

To complete our equation, we need the intercept, b.

We take advantage of the fact that the point corresponding to the two means, namely, $(\bar{x}, \bar{y})$, is exactly on the line. We won't prove this just yet.

The point $(2, 8/3)$ is on the line, so it satisfies the equation $y = mx + b$. Rearranging

$$b = y - mx$$

And in this case, these are means so we can write

$$b = \bar{y} - m\bar{x}$$

$$= 8/3 - (1 \cdot 2) = 2/3$$

And that's how the calculation is done. We have the same answer as Desmos, which is reassuring.

Suppose we start with data that has already been centered. What value will we get for b?

At this point, you might want to break out a spreadsheet app like Excel or Numbers or the one on Google Drive, and try to code the calculations we just did.

If you're adventurous, you could make a new spreadsheet for the gasoline example and see if you can get the same numbers for m and b of the regression equation as Desmos gives.

My version is at this link on Google Drive

`https://docs.google.com/spreadsheets/d/1umQvka2ECjHB6rKcInvYP0qCnUvJ-A0`
`edit?usp=sharing`

It looks like this:

	A	B	C	D	E	F	G
1	2002	-4.5	20.25		0.7	-0.945	4.2525
2	2003	-3.5	12.25		1.04	-0.605	2.1175
3	2004	-2.5	6.25		1.15	-0.495	1.2375
4	2005	-1.5	2.25		1.38	-0.265	0.3975
5	2006	-0.5	0.25		1.86	0.215	-0.1075
6	2007	0.5	0.25		1.7	0.055	0.0275
7	2008	1.5	2.25		2.55	0.905	1.3575
8	2009	2.5	6.25		1.29	-0.355	-0.8875
9	2010	3.5	12.25		2.22	0.575	2.0125
10	2011	4.5	20.25		2.56	0.915	4.1175
11							
12	2006.5				1.645		
13			82.5				14.525
14							
15	m =	0.176					
16	b =	-351.621					

Column C is $(x_i - \bar{x})^2$ and the sum is in cell C13. Column G is the same for $(x - \bar{x})(y - \bar{y})$. The ratio of the sum of the second to the sum of the first is m.

You can see that the values for m and b match, if you enter the same data on Desmos.

correlation statistic

If you notice, the linear regression output from Desmos also lists something called r and r^2. These quantify how good the correlation is. The closer r is to 1, the better.

We need SS_{xy} and SS_{xx} from before.

We also calculate SS_{yy}. It should be clear how to do this.

Then

$$r = \cdot \frac{SS_{xy}}{\sqrt{SS_{xx}} \cdot \sqrt{SS_{yy}}}$$

For the toy example, I get

$$SS_{yy} = (-2/3)^2 + (-2/3)^2 + (4/3)^2 = \frac{24}{9} = \frac{8}{3}$$

Then, the calculation of r is

$$r = \frac{2}{\sqrt{2} \cdot \sqrt{8/3}} = \sqrt{\frac{3}{4}} = 0.866$$

And for the gasoline example, I get $r = 0.827$ (see the spreadsheet).