

Paired Data(bivariate data) and Scatterplots:

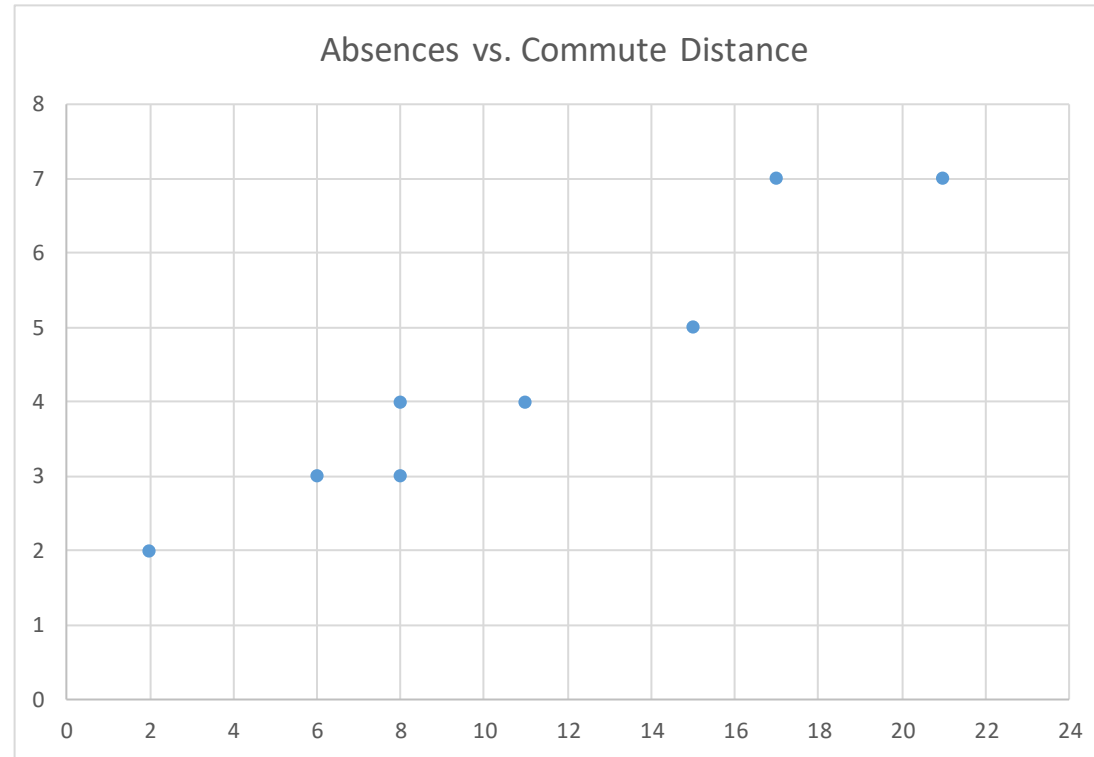
When data consists of pairs of values, it's sometimes useful to plot them as points called a scatterplot.

A company recorded the commuting distance in miles and number of absences in days for a group of its employees over the course of a year.

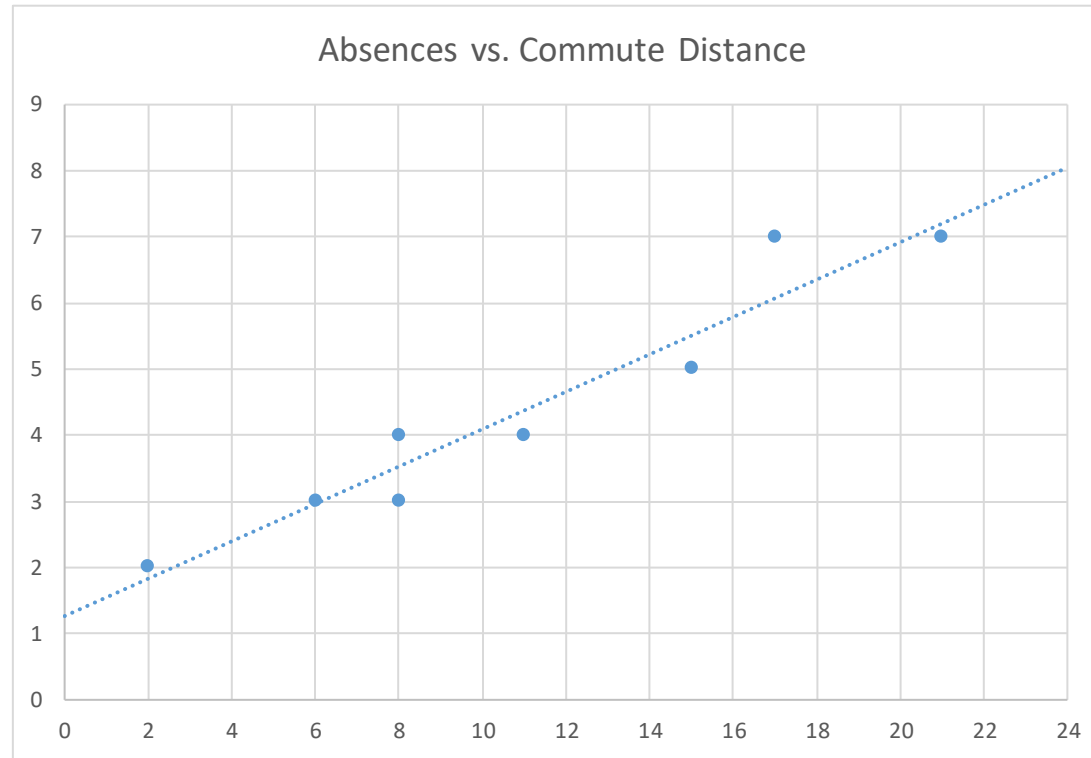
Commuting Distance	Number of Absences
8	4
21	7
6	3
8	3
2	2
15	5
17	7
11	4



Here's the scatterplot of number of absences vs. commuting distance.



Scatterplots can help reveal relationships between the variables being measured in the paired data. The most commonly searched for relationship is a linear relationship. In this example, the data values do appear to cluster on a line.



When the points cluster on a line, the variables being measured are said to be linearly correlated. If the line they cluster on has a positive slope, then the variables are said to be positively correlated.

If the cluster line has a negative slope, then the variables are said to be negatively correlated.

If the points don't cluster on a non-vertical, non-horizontal line, then the variables are said to be uncorrelated.

When variables are positively correlated, larger values of one variable are associated with larger values of the other variable.

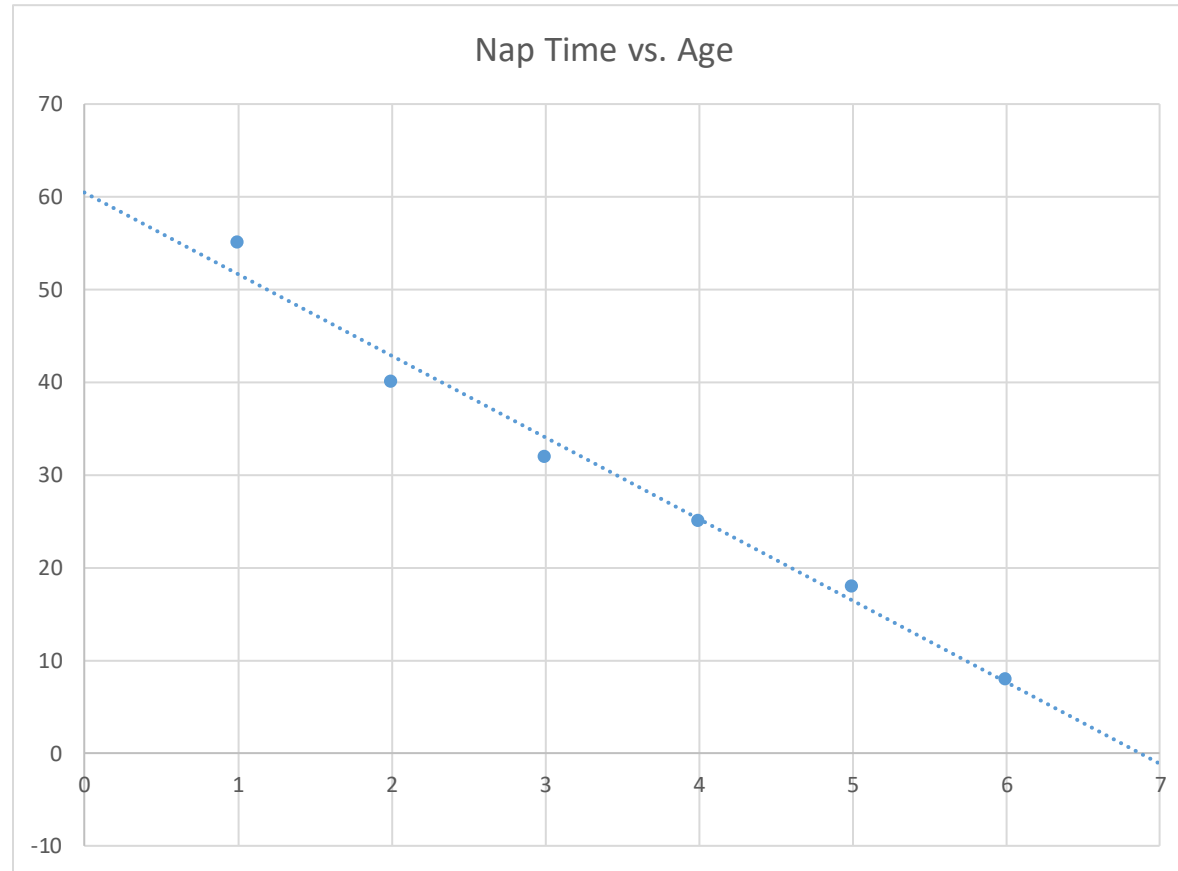
When variables are negatively correlated, larger values of one variable are associated with smaller values of the other variable.

Here is a paired data set of Nap time in minutes along with Age of child in years.

Age(years)	Nap time(minutes)
1	55
2	40
3	32
4	25
5	18
6	8



Here's its scatterplot:



Nap time and age appear to be negatively(linearly) correlated for this group of children, since the data values do appear to cluster on a line with a negative slope.

There is a quantity for determining the strength of a linear relationship as well as its direction. It's called the correlation coefficient.

$$\text{correlation coefficient} = r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}} = \frac{\sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)}{n - 1}$$

r is between -1 and 1, inclusive.



A value of -1 indicates a perfect negative correlation, i.e. all the points lie on a line with a negative slope.

A value of 1 indicates a perfect positive correlation, i.e. all the points lie on a line with a positive slope.

A value close to 0 means the variables are uncorrelated, i.e. the points do not cluster on a non-vertical, non-horizontal line.

The closer to 1, the stronger the positive correlation.

The closer to -1, the stronger the negative correlation.

The closer to zero, the weaker the correlation.

It's possible for the correlation coefficient to be undefined. For example, if all the y-coordinates are the same, it's undefined.

Let's calculate the correlation coefficients of the previous data sets using the following table:

	x	y	xy	x^2	y^2
1	8	4	32	64	16
2	21	7	147	441	49
3	6	3	18	36	9
4	8	3	24	64	9
5	2	2	4	4	4
6	15	5	75	225	25
7	17	7	119	289	49
8	11	4	44	121	16
Total	88	35	463	1244	177

$$\sum x$$

$$\sum y$$

$$\sum xy$$

$$\sum x^2$$

$$\sum y^2$$

Plugging the table values into the formula $r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$, leads to

$$r = \frac{463 - \frac{88 \cdot 35}{8}}{\sqrt{1244 - \frac{(88)^2}{8}} \sqrt{177 - \frac{(35)^2}{8}}} = \frac{78}{\sqrt{276} \sqrt{23.875}} = .960878...$$

This value indicates a strong positive correlation.

	x	y	xy	x^2	y^2
1	1	55	55	1	3025
2	2	40	80	4	1600
3	3	32	96	9	1024
4	4	25	100	16	625
5	5	18	90	25	324
6	6	8	48	36	64
Total	21	178	469	91	6662

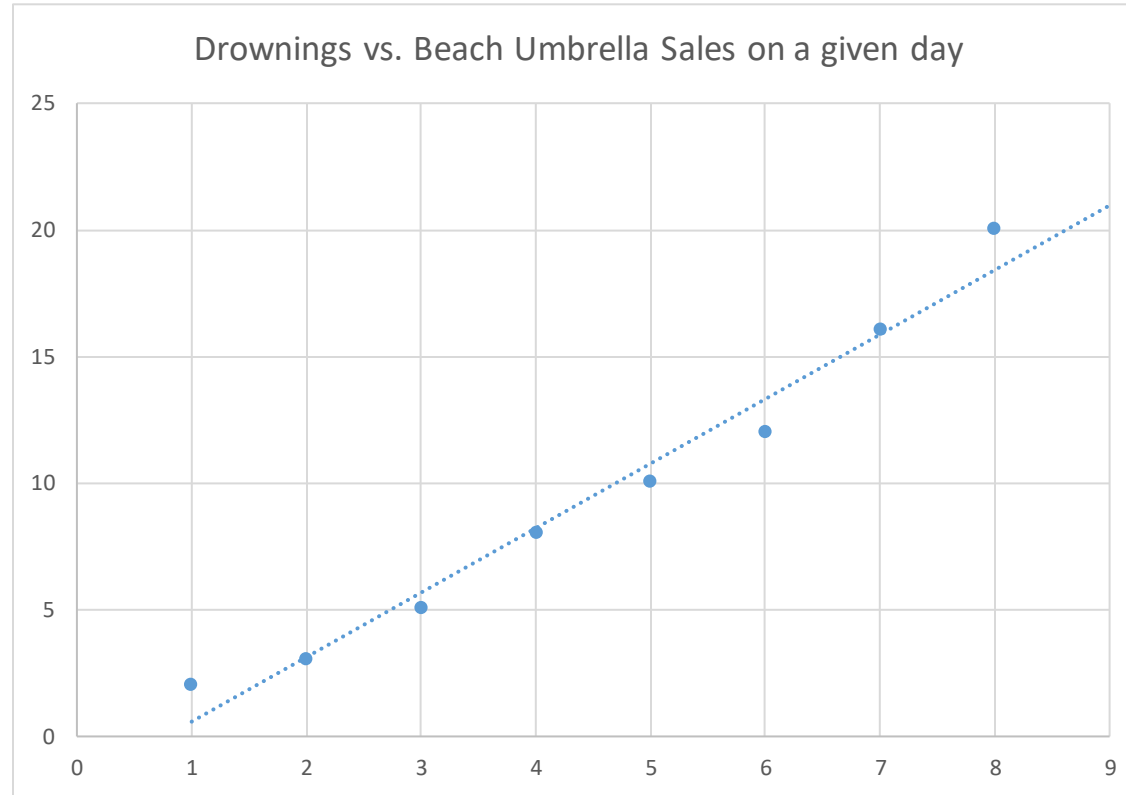
$$\sum x \quad \sum y \quad \sum xy \quad \sum x^2 \quad \sum y^2$$

Plugging the table values into the formula $r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$, leads to

$$r = \frac{469 - \frac{21 \cdot 178}{6}}{\sqrt{91 - \frac{(21)^2}{6}} \sqrt{6662 - \frac{(178)^2}{6}}} = \frac{-154}{\sqrt{17.5} \sqrt{1381.3}} = -.990495...$$

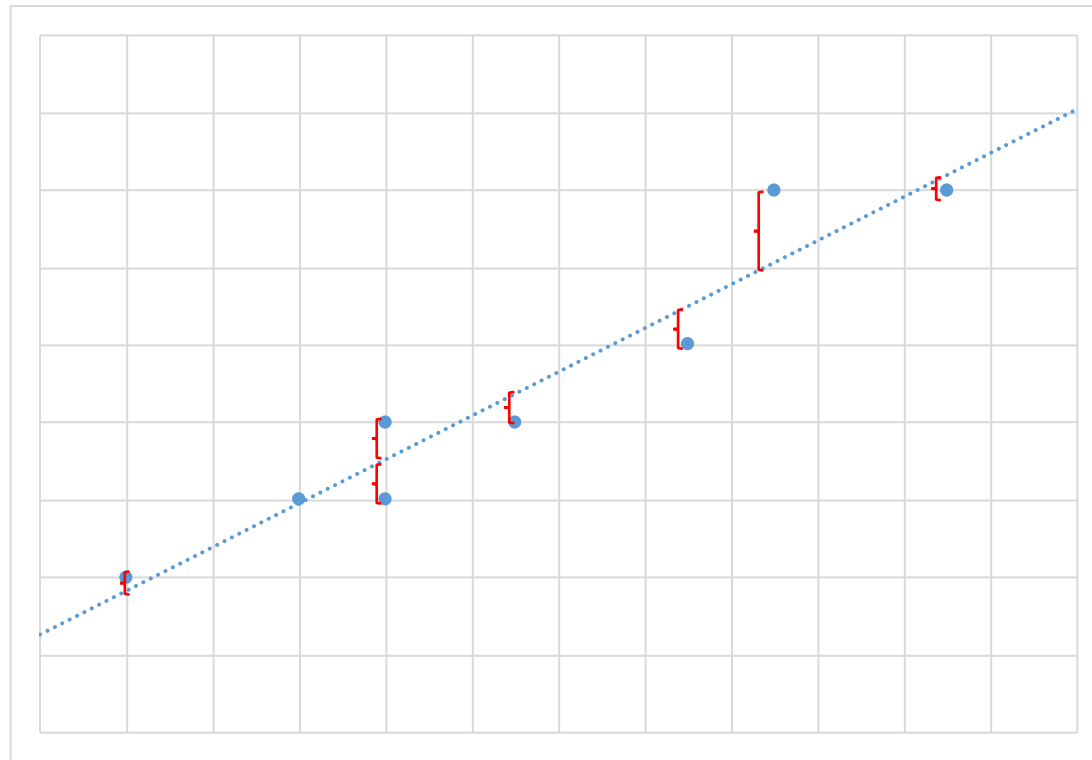
This value indicates a strong negative correlation.

Just because variables are correlated doesn't mean that they are causally related!



I wouldn't conclude that beach umbrellas cause drowning, but both are correlated to temperature/season.

The official name of a line of best fit(cluster line) is regression line, and there are several ways of calculating a regression line. The most popular method chooses the line which minimizes the sum of the squares of the vertical deviations from the points in the scatterplot.



It's called the least squares regression line.

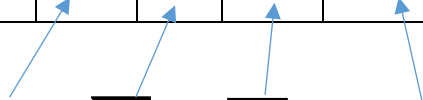
The equation of the least squares regression line is $\hat{y} = b_1x + b_0$, with

$$b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = r \frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}.$$



Let's find the least squares regression equations for the previous sets of paired data.

	x	y	xy	x^2
1	8	4	32	64
2	21	7	147	441
3	6	3	18	36
4	8	3	24	64
5	2	2	4	4
6	15	5	75	225
7	17	7	119	289
8	11	4	44	121
Total	88	35	463	1244

$$\sum_x \quad \sum_y \quad \sum_{xy} \quad \sum_{x^2}$$


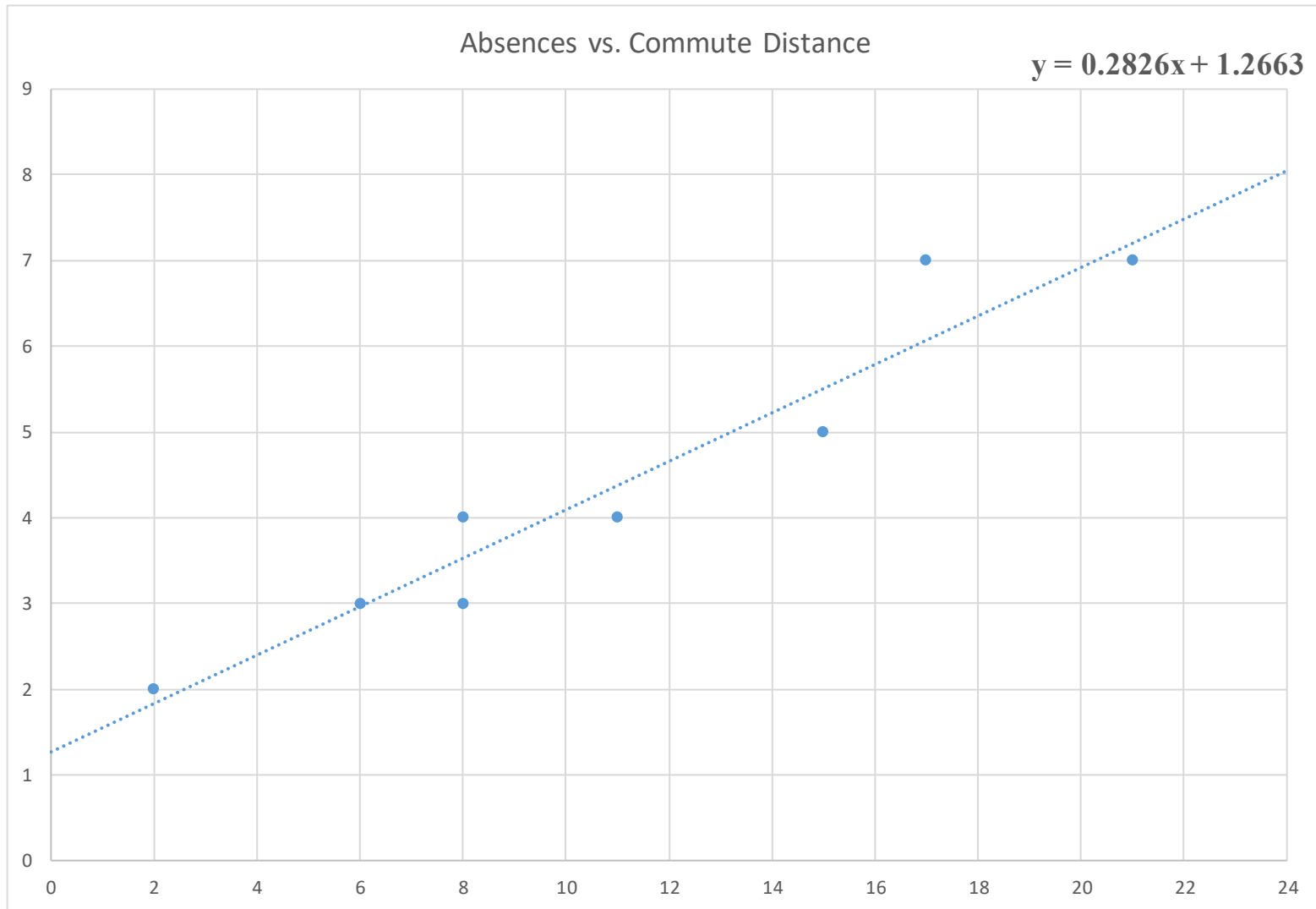
Plugging the table values into the formulas $b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$ and $b_0 = \bar{y} - b_1\bar{x}$,

leads to

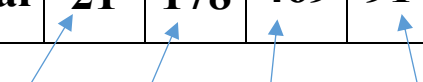
$$b_1 = \frac{463 - \frac{88 \cdot 35}{8}}{1244 - \frac{(88)^2}{8}} = \frac{78}{276} = .28260... \text{ and } b_0 = \frac{35}{8} - \frac{78}{276} \cdot \frac{88}{8} = \frac{233}{184} = 1.2663...$$

$$\hat{y} = .2826x + 1.2663$$

It's equation can also be determined using statistical software, like Excel.



	x	y	xy	x^2
1	1	55	55	1
2	2	40	80	4
3	3	32	96	9
4	4	25	100	16
5	5	18	90	25
6	6	8	48	36
Total	21	178	469	91



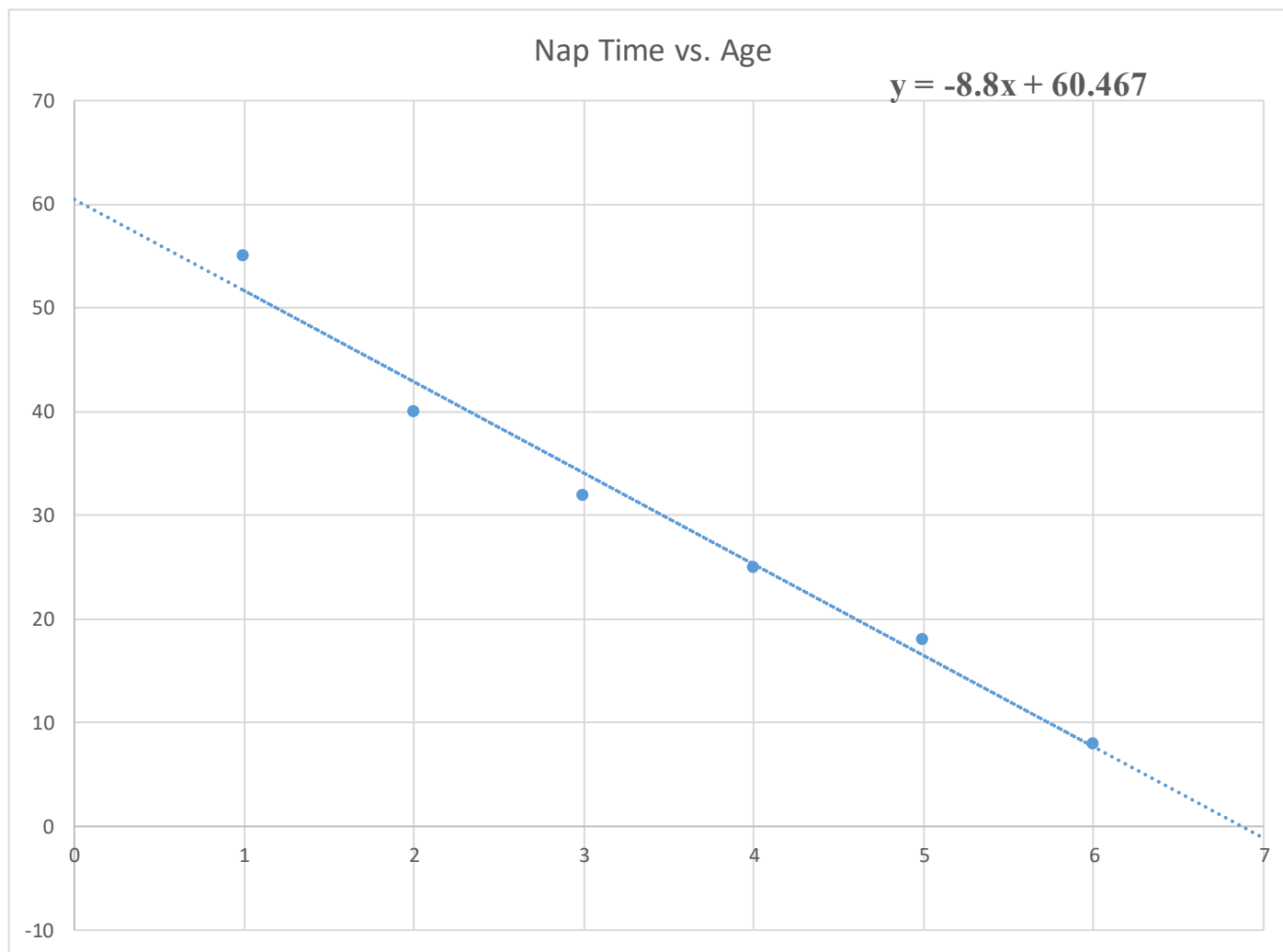
$$\sum x \quad \sum y \quad \sum y \quad \sum x^2$$

Plugging the table values into the formulas $b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$ and $b_0 = \bar{y} - b_1 \bar{x}$,

leads to

$$b_1 = \frac{469 - \frac{21 \cdot 178}{6}}{91 - \frac{(21)^2}{6}} = \frac{-154}{17.5} = -8.8 \text{ and } b_0 = \frac{178}{6} - (-8.8) \frac{21}{6} = 60.4666...$$

$$\hat{y} = -8.8x + 60.467$$



Sometimes the equation of a regression line or its graph is used to make predictions about a value of the variables that wasn't measured, i.e. wasn't a value in the data set.

Use the least squares regression line equation, $\hat{y} = .2826x + 1.2663$ to answer the following:

What's the predicted number of absences for an employee with a 10 mile commute?

$$y = .2826(10) + 1.2663 = 4.0923 \approx 4\text{days} \text{ (interpolation, safe)}$$

This prediction is called an interpolation because the 10 miles falls within the distance values recorded in the data set. Interpolation predictions are considered to be safe and generally reliable.

What's the predicted number of absences for an employee with a 24 mile commute?

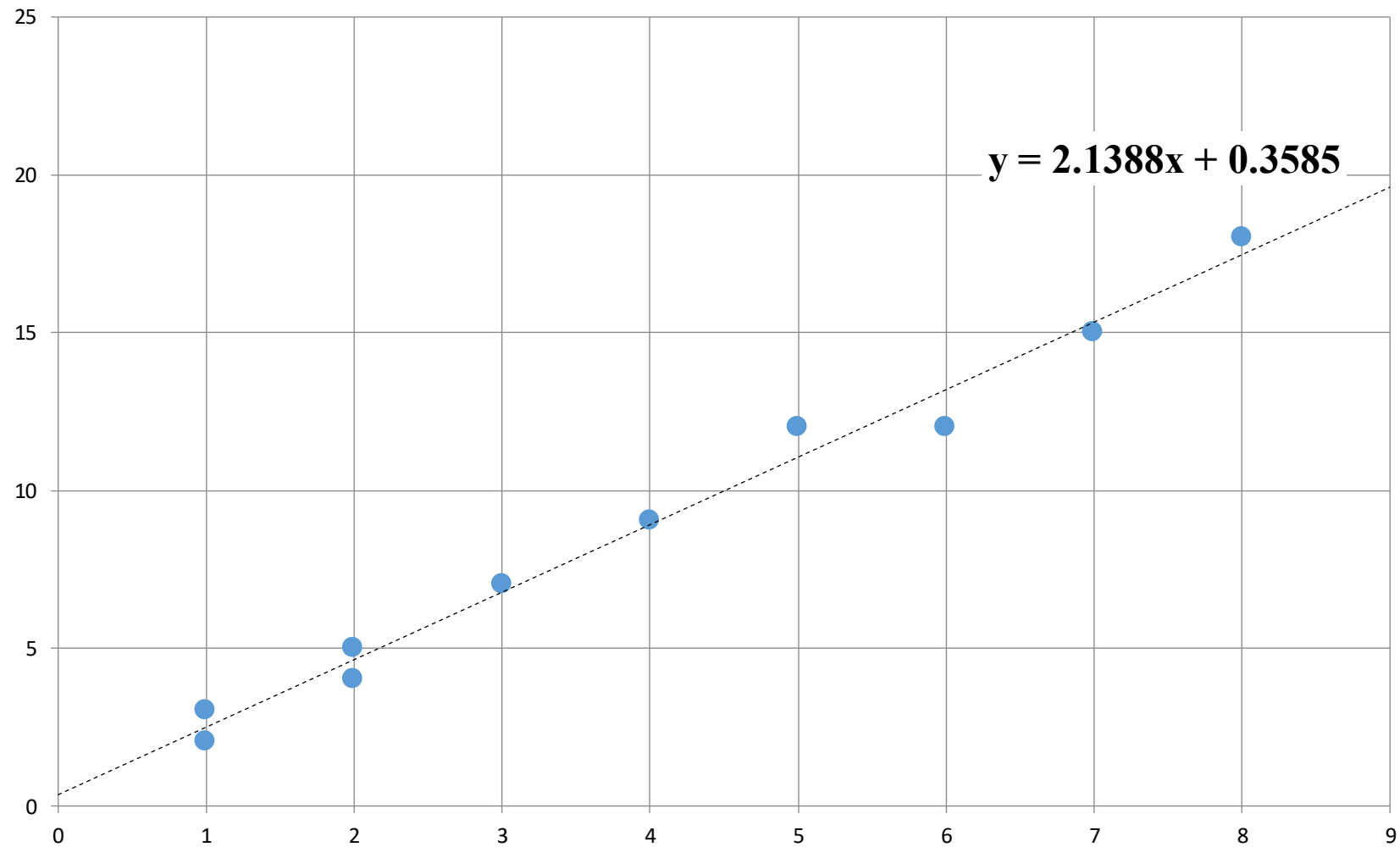
$$y = .2826(24) + 1.2663 = 8.0487 \approx 8\text{days} \text{ (extrapolation, dangerous)}$$

This prediction is called an extrapolation because the 24 miles falls outside of the distance values recorded in the data set. Extrapolation predictions are considered to be risky and generally unreliable.

Here's a set of paired data.

X	Y
1	3
2	5
2	4
1	2
4	9
5	12
6	12
3	7
7	15
8	18

Here's its scatterplot, correlation coefficient, and regression line from Excel.



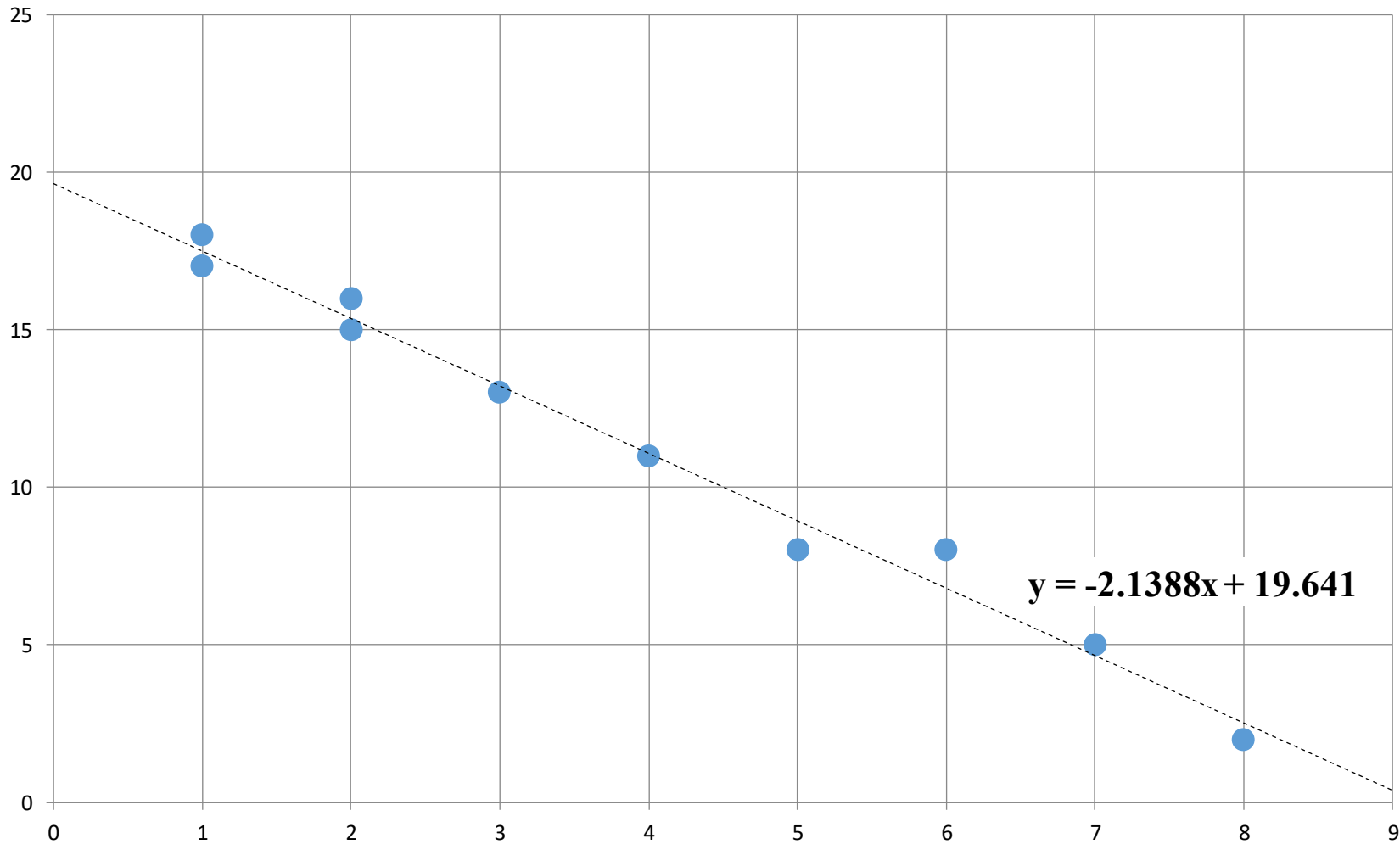
Correlation Coefficient

0.992773656

Here's a set of paired data.

X	Y
1	17
2	15
2	16
1	18
4	11
5	8
6	8
3	13
7	5
8	2

Here's its scatterplot, correlation coefficient, and regression line from Excel.

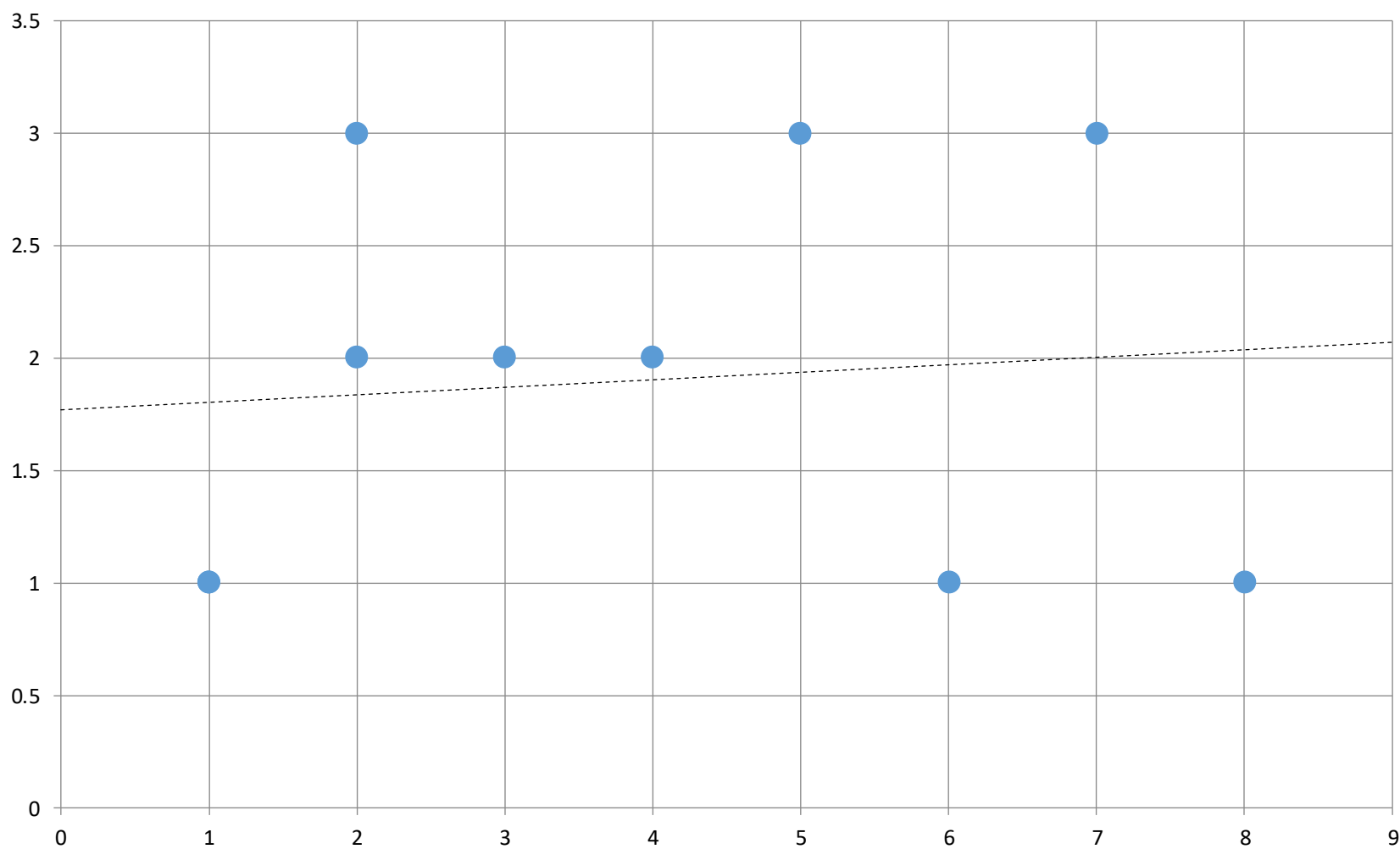


Correlation Coefficient
-0.992773656

Here's a set of paired data.

X	Y
1	1
2	2
2	3
1	1
4	2
5	3
6	1
3	2
7	3
8	1

Here's its scatterplot, correlation coefficient, and regression line from Excel.

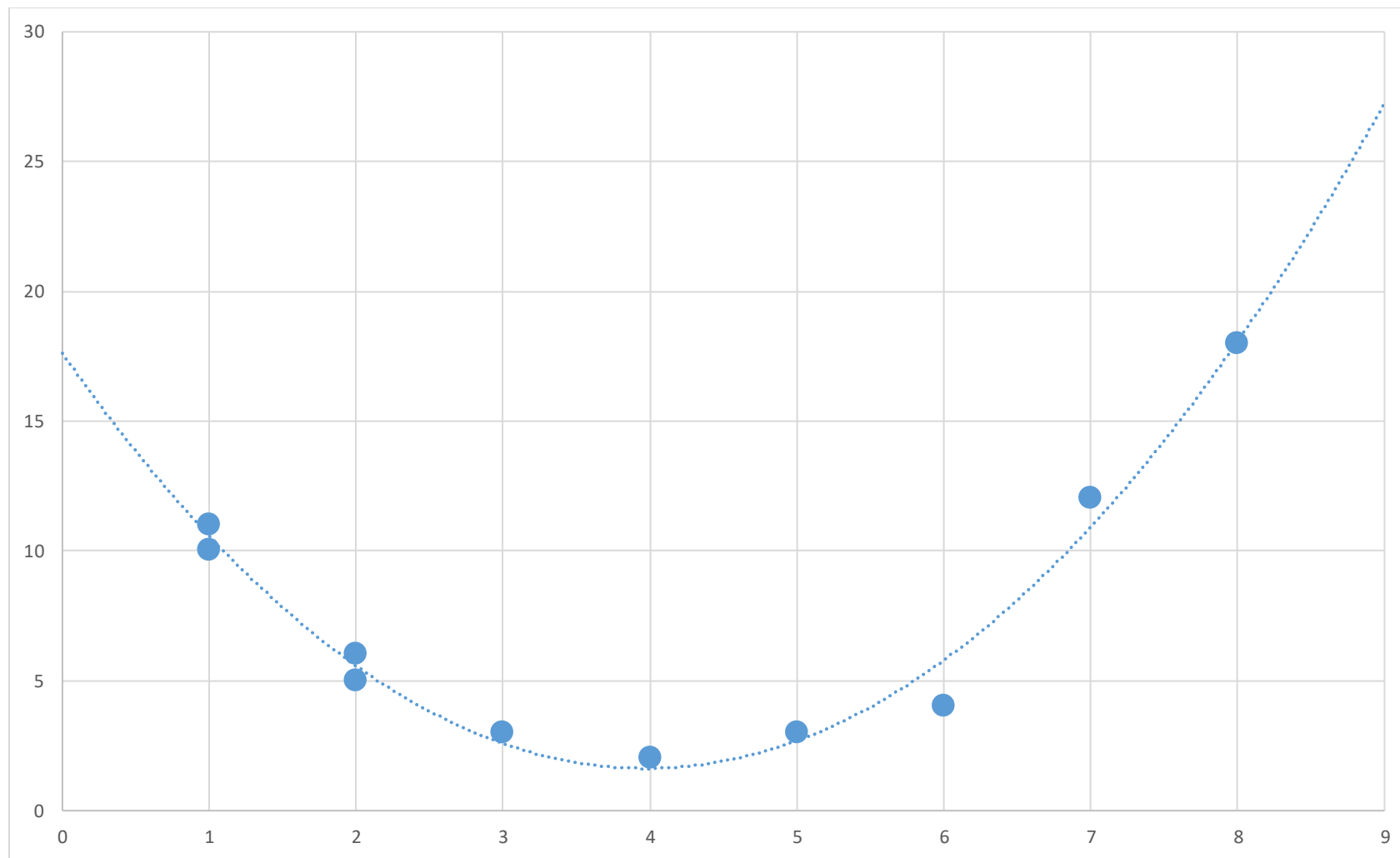


Correlation Coefficient
0.095889913

Here's a set of paired data.

X	Y
1	11
2	6
2	5
1	10
4	2
5	3
6	4
3	3
7	12
8	18

Here's its scatterplot and correlation coefficient from Excel.



Correlation Coefficient

0.328327953

For a paired data set, $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, without a regression line, the standard predictor of a y -value for a given x -value is the mean of the y -values, $\bar{y}$. The amount of variation that the y -values have from their mean value, $\bar{y}$, is measured by summing all the squared deviations, $\sum (y - \bar{y})^2$, and it's called the total variation. When the least squares regression line is used, the total variation can be split into two pieces and related to the total variation by the equation:

$$\underbrace{\sum (y - \bar{y})^2}_{\text{Total Variation}} = \underbrace{\sum (y - \hat{y})^2}_{\text{Unexplained Variation}} + \underbrace{\sum (\hat{y} - \bar{y})^2}_{\text{Explained Variation}}.$$

The closer the least squares regression line is to the data points, the closer the Unexplained Variation is to zero. The more that the regression line is like the horizontal line $y = \bar{y}$, the closer the Explained Variation is to zero.

The proportion of the variation in the y -values that is explained by the least squares regression line is called the coefficient of determination, R^2 , and it's frequently expressed as a percent.

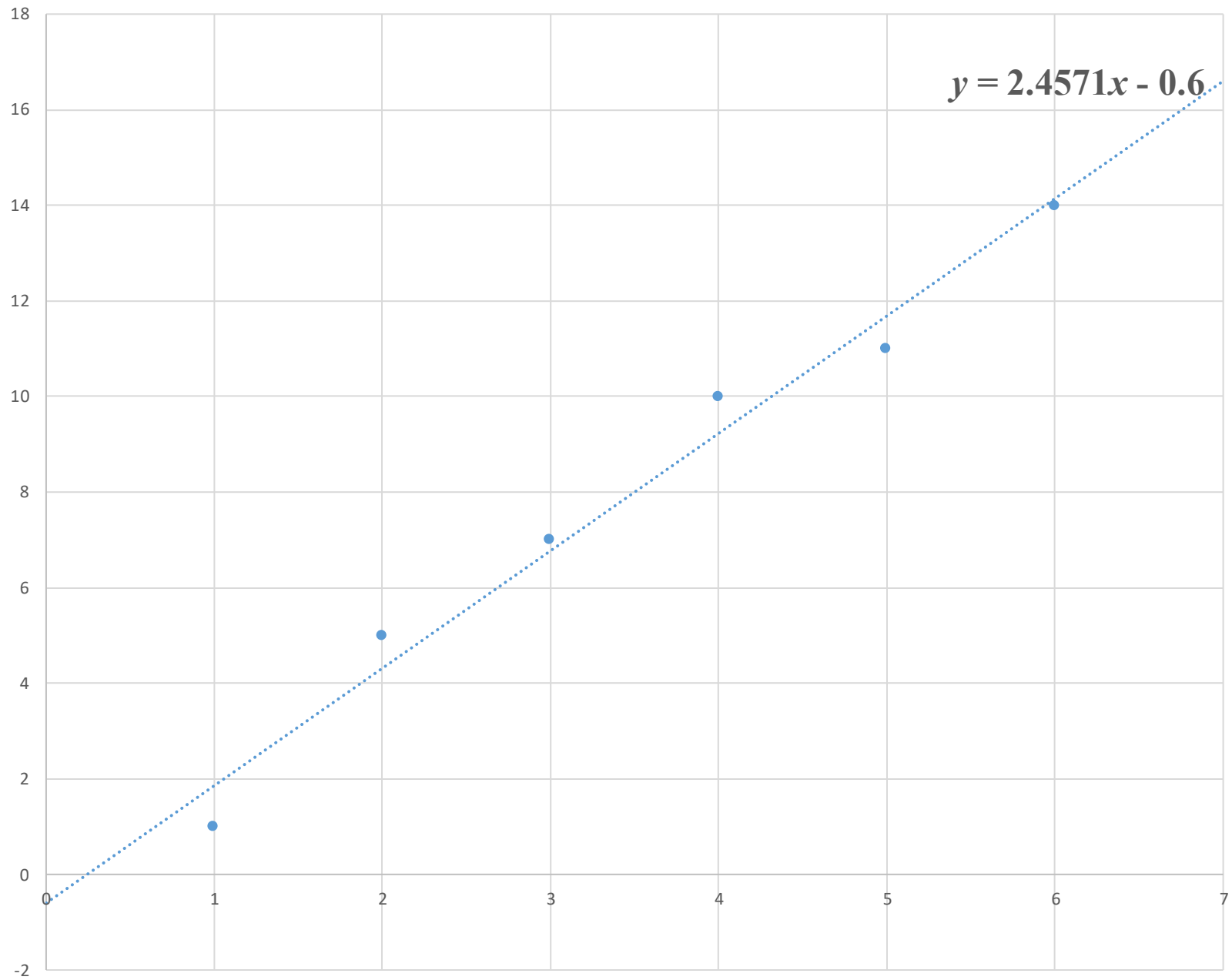
$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

It's interpreted as the percent of the variation in the y -values that's explained by the Least Squares regression line.

Example:

x	y
1	1
2	5
3	7
4	10
5	11
6	14

Scatterplot



x	y	$\hat{y}$	$(\hat{y} - \bar{y})^2$	$(y - \hat{y})^2$	$(y - \bar{y})^2$
1	1	1.8571	37.7347	0.7347	49
2	5	4.3143	13.5845	0.4702	9
3	7	6.7714	1.5094	0.0522	1
4	10	9.2286	1.5094	0.5951	4
5	11	11.6857	13.5845	0.4702	9
6	14	14.1429	37.7347	0.0204	36
Total			105.6571	2.3429	108
			Explained Variation	Unexplained Variation	Total Variation

Notice that the sum of the Explained Variation and the Unexplained Variation is equal to the Total Variation.

So the coefficient of determination is $\frac{105.6571}{108} = .978306... = 97.8306...\%$. This means that the least squares regression line explains about 98% of the variation in the y -values.

The coefficient of determination, R^2 , will always have a value between 0(0%) and 1(100%), inclusive. Values close to 0 or 0% indicate that the least squares regression line does a poor job of predicting the y -values. Values close to 1 or 100% indicate that the least squares regression line does a good job of predicting the y -values.

There is a connection between the coefficient of determination and the correlation coefficient: $R^2 = r^2$. In other words, the square of the correlation coefficient is equal to the coefficient of determination.

Example:

Suppose that the correlation coefficient for a set of paired data is $r = .91$. Find the value of the coefficient of determination, and interpret its value.

$R^2 = (.91)^2 = .8281$, so the least squares regression line explains 82.81% of the variability in the y -value.