

How to measure association I: Correlation.

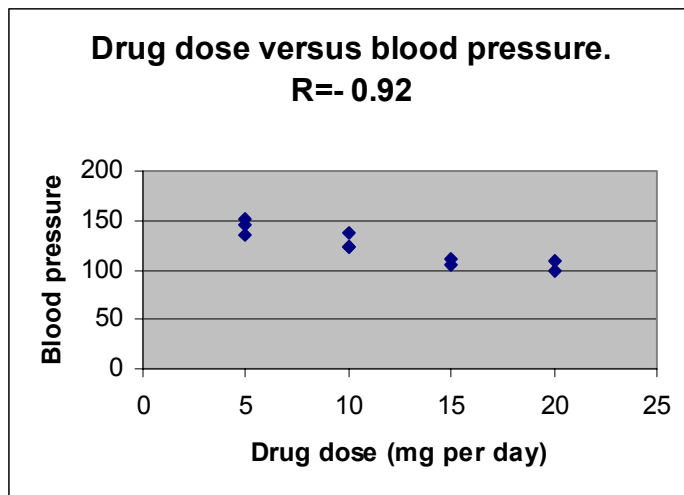
1. Measuring association using correlation and regression

We often would like to know how one variable, such as a mother's weight, is related to another variable, such as a baby's birthweight. We might be interested in the relationship between a patient's blood pressure and the amount of drug the patient takes per day.

Suppose that we have data on blood pressure and drug dose such as in Table <BP-drug dose>.

Table <BP-drug dose>.

Drug dose (mg per day)	Blood pressure
5	151
5	145
5	136
10	137
10	124
10	124
15	111
15	105
20	110
20	98



Two questions we may ask are:

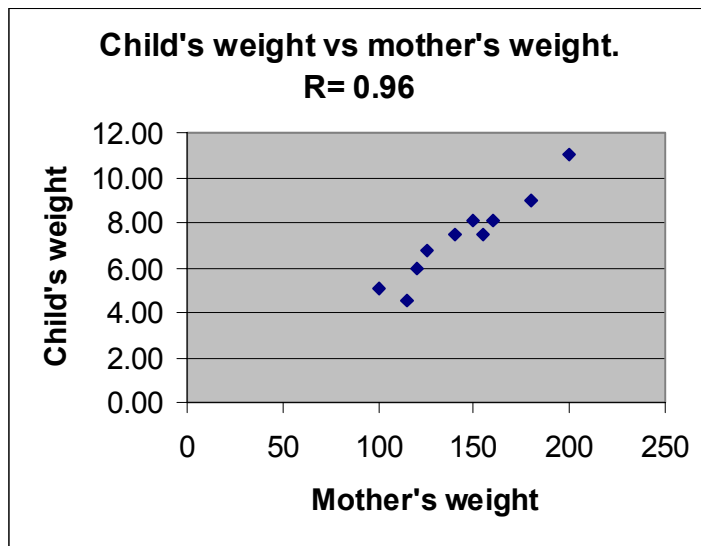
1. If I know how much drug the patient was given, how well can I predict their blood pressure? Put another way, we can ask how much of the variability in blood pressure can be explained by differences in the amount of drug the patient takes. To answer this question, we use correlation, which we discuss in this chapter.

2. For a unit change in the amount of drug given, how much change in blood pressure do we expect? To answer this question, we use regression, which we discuss in the next chapter.

As another example of where we use correlation or regression, suppose that we are interested in babies who are born with low birthweights, and want to examine factors that affect birthweight. We might have data on mother's weight and baby's birthweight as in Table <Birthweights>.

Table <Birthweights>.

Mother's weight	Child's birthweight
100	5.10
115	4.50
120	6.00
125	6.80
140	7.50
150	8.10
155	7.50
160	8.10
180	9.00
200	11.00



Again, two questions we may ask are:

1. If I know the mother's weight, how well can I predict the baby's weight? Put another way, we can ask how much of the variability in baby's weight can be explained by differences in the mother's weight. To answer this question, we use correlation.

2. For a unit change in the mother's weight (one pound increase), how much change in baby's weight do we expect? To answer this question, we use regression.

You may notice that all the variables we are considering (blood pressure, weight, dose) are measured on a continuous scale, and these are suitable for correlation and regression. If we want to measure association between categorical variables (such as male/female, Republican/Democrat, pass/fail, yes/no, and so on) we use statistics such as the chi-square test which we'll look at in a later chapter.

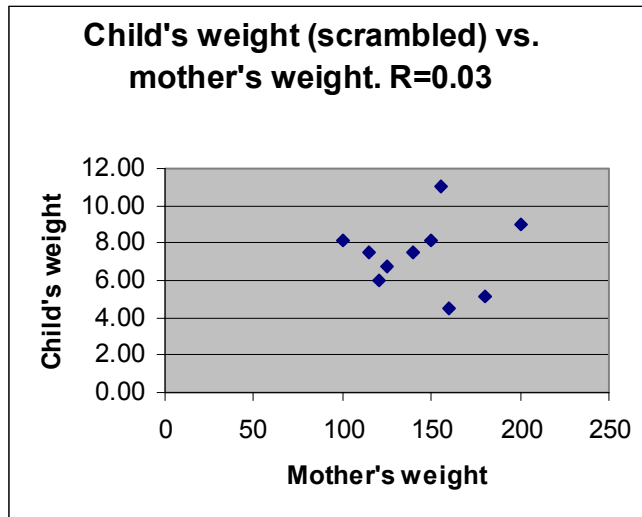
We are going to focus mainly on the most widely used correlation measure, which is R , the Pearson linear correlation coefficient. Later on, we'll look at another correlation measure, the Spearman rank correlation coefficient, which is sometimes better to use than the Pearson measure.

2. Correlation can be positive, zero, or negative (ranging from 1.0 to -1.0)

Correlation can be positive as in the birthweight example or negative as in the drug/blood pressure example. By definition, using the formula we'll see in the next section, the maximum (positive) correlation is 1.0. In the birthweight example, correlation was nearly perfect at $R = 0.96$. The minimum possible (negative) correlation is -1.0. In the drug versus blood pressure example, correlation was strongly negative with $R = -0.92$. Correlation can also be near zero, as shown in Table <Scrambled birthweights>, where we have scrambled the children's birthweights, and see $R = 0.03$.

Table <Scrambled birthweights>

Mother's weight	Child's weight
100	8.10
115	7.50
120	6.00
125	6.80
140	7.50
150	8.10
155	11.00
160	4.50
180	5.10
200	9.00



3. How to calculate the Pearson linear correlation coefficient

We'll first define the Pearson linear correlation coefficient, and then look at how to interpret it.

Recall the formula for variance from the chapter on descriptive statistics. Variance describes variability around the mean value.

$$\text{Variance} = \frac{\sum_i (x_i - \bar{x})^2}{N}$$

Covariance has a formula similar to that for the variance.

$$\text{Covariance}(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{N}$$

Correlation uses the covariance of two variables. The correlation of two variables, x and y, is equal to the covariance of x and y divided by a number that makes correlation be between -1.0 and 1.0.

$$\text{Correlation}(x, y) = R = \frac{\text{Covariance}(x, y)}{\sqrt{\text{Var}(x) * \text{Var}(y)}}$$

The term in the denominator, the square root of $\text{Var}(x) * \text{Var}(y)$, just forces the correlation coefficient to be between -1.0 and 1.0; it doesn't affect how we interpret the correlation coefficient, so we won't look at it any further.

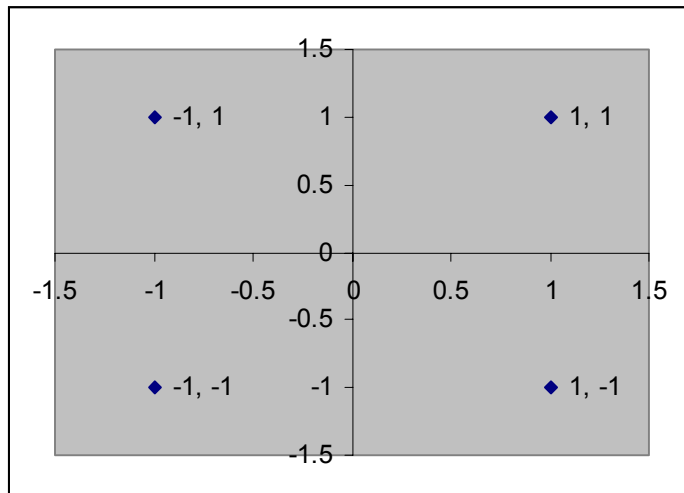
4. How to interpret the correlation coefficient

Let's look at what the correlation coefficient tells us. We'll start with just four points, one from each quadrant, as shown in Table <Points in 4 quadrants>. Quadrant 1 is labeled here as (1,1), quadrant 2 is labeled (-1,1), quadrant 3 is labeled (1, -1), and quadrant 4 is labeled (-1, -1). For any data set, we can force the mean to be at (0,0) by subtracting the mean of all the x values from the x value for each point and the mean of all the y values from the y value for each point. For these "Mean corrected" values, the mean is now at (0,0), and every point must fall into one of the four quadrants relative to the mean.

Table <Points in 4 quadrants>.

x value	y value
1	1
-1	1
1	-1
-1	-1

Figure <Points in 4 quadrants>.



Now, let's look again at the formula for covariance.

$$\text{Covariance}(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{N}$$

We've specified that we subtract the means, so the new mean value of x is zero and the new mean value of y is 0, and the formula for covariance then simplifies as follows.

$$\text{Covariance}(x,y) = \frac{\sum_i (x_i)(y_i)}{N}$$

Consider a point in quadrant 1 in Figure <Points in 4 quadrants>, such as the point (1,1). In the formula for covariance, we put the point (1,1), into the term $(x_i)(y_i)$, and we get $1*1 = 1$, which is a positive number. For the term $(x_i)(y_i)$, every point in quadrant 1 will give a positive value, because we are multiplying two positive numbers.

Next, consider a point in quadrant 3, such as (-1,-1). In the formula for covariance, we put the point (-1,-1) into the term $(x_i)(y_i)$, which gives us $-1*-1 = 1$, which is again a positive number. For the term $(x_i)(y_i)$, every point in quadrant 3, where we are multiplying two negative numbers, which will give a positive value.

Points in quadrants 2 and 4 will give us negative values for the term $(x_i)(y_i)$. In quadrant 2, we see that $-1* 1 = -1$, and in quadrant 4, we see that $-1* 1 = -1$.

If all the points in our data set fall into quadrant 1 or quadrant 3 with respect to the mean, then every point will contribute a positive value to the covariance, which will in turn give us a large positive correlation.

In contrast, if all the points in our data set fall into quadrant 2 or quadrant 4 with respect to the mean, then every point will contribute a negative value to the covariance, which will in turn give us a large negative correlation.

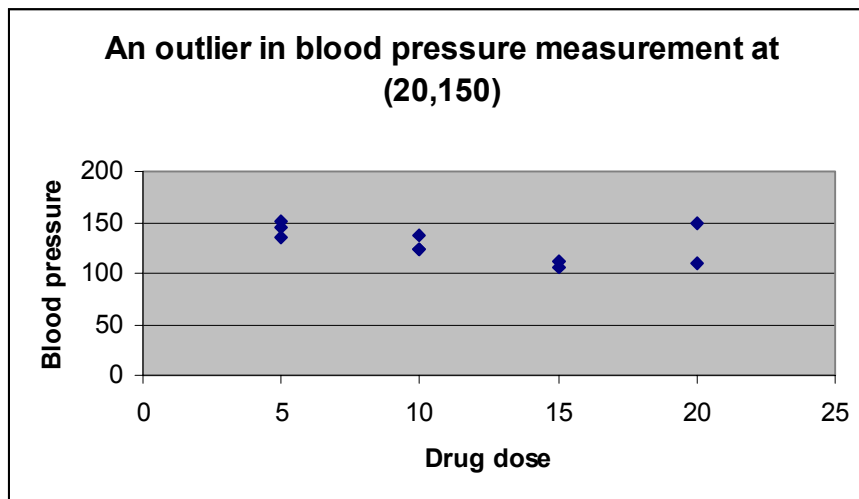
If points are scattered across all four quadrants, we will get a mixture of positive and negative terms that tend to cancel each other out, giving a correlation near zero.

5. Potential problems with Pearson linear correlation

The Pearson linear correlation coefficient can be greatly affected by a single observation. In particular, a single point (an outlier) that falls a long way from other points in the x-y plane can greatly increase or decrease the Pearson R. For example, let's look again at the data on drug dose versus blood pressure, but suppose that the last patient, instead of having a blood pressure measurement of 98, has a value of 150 as in Table <Outlier in BP-drug dose>. For these data, the Pearson correlation coefficient is $R = -0.47$, which is a large change from the $R = -0.92$ we had before changing this single point. When we see an individual point that is so influential in determining the value of our statistic, we should consider the possibility that there was an error in the measurement, and make sure that we are not being misled.

Table <Outlier in BP-drug dose>.

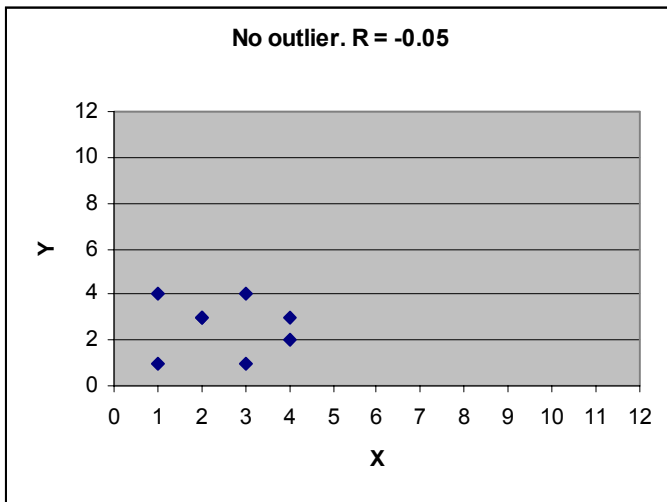
Drug dose (mg per day)	Blood pressure
5	151
5	145
5	136
10	137
10	124
10	124
15	111
15	105
20	110
20	150



A single outlier can also make a weak correlation appear much stronger. For the data in <Table no-outlier>, the correlation coefficient is quite small, $R = -0.05$.

<Table no-outlier>

x value	y value
1	4
1	1
2	3
2	3
3	1
3	4
4	3
4	2

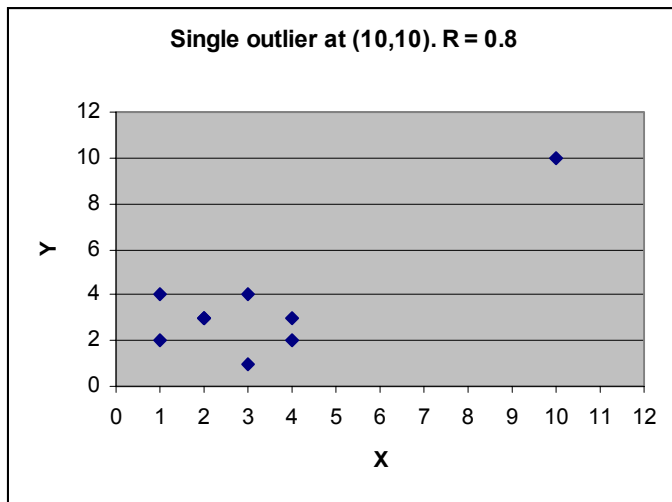


However, if we add a single observation at (10, 10), as shown in Table <Table Single-outlier> and Figure <Figure Single-outlier>, we change the correlation coefficient from $R = -0.05$ to $R = 0.81$.

<Table Single-outlier>

x value	y value
1	4
1	2
2	3
2	3
3	1
3	4
4	3
4	2
10	10

<Figure Single-outlier>

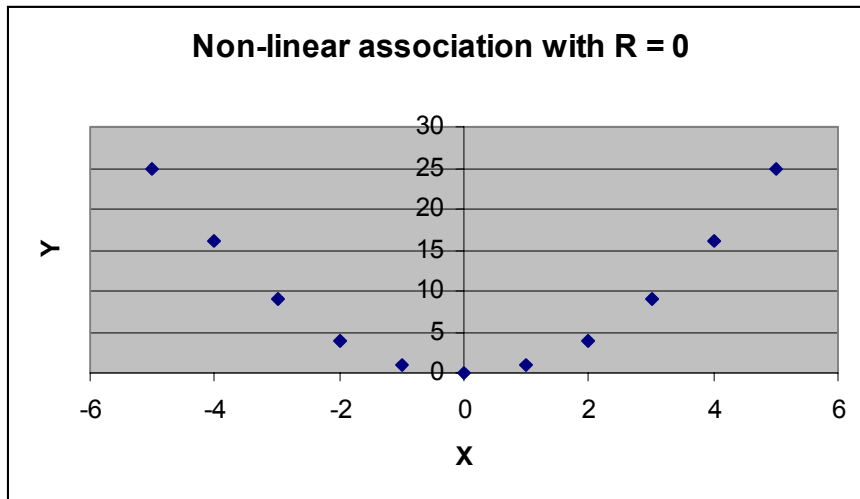


So we see that the Pearson linear correlation coefficient may be very sensitive to a single point. In such situations, we may choose to use an alternative association measure, the Spearman rank correlation coefficient, which we'll look at shortly.

The Pearson linear correlation coefficient is not good at detecting genuine but non-linear associations between variable. Suppose that we have values such as those in Table <Table non-linear relation> and Figure < Figure non-linear relation>. Although there is clearly a relationship between x and y, the correlation coefficient is $R = 0.0$. This example shows that it is always a good idea to graph your data, and not to rely completely on a statistic.

<Table non-linear relation>.

x value	y value
-5	25
-4	16
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9
4	16
5	25



6. Spearman rank correlation: an alternative to Pearson correlation

We saw that the Pearson correlation coefficient may be greatly affected by single influential points (outliers). Sometimes we would like to have a measure of association that is not so sensitive to single points, and at those times we can use Spearman rank correlation.

Recall that, when we calculate the mean of a set of numbers, a single extreme value can greatly increase the mean. But when we calculate the median, which is based on ranks, extreme values have very little influence. The same idea applies to Pearson and Spearman correlation. Pearson uses the actual values of the observations, while Spearman uses only the ranks of the observations, and thus, like the median, is not much affected by outliers.

Most statistics packages will calculate either Pearson or Spearman, but Excel will only do Pearson. The easiest way to get Spearman is to replace each observation by the rank value of each observation, and then calculate the Pearson coefficient using the ranks.

For the outlier examples, recall that the Pearson correlation is $R = -0.05$ excluding the outlier and $R = 0.81$ including the outlier.

For these data, the Spearman rank correlation is $R_s = -0.10$ excluding the outlier and $R_s = 0.24$ including the outlier. Let's do the calculations. Here's the data excluding the single outlier. I've assigned the rank to each value, with ties given the average rank.

x value	x rank	y value	y rank
1	1.5	4	7.5
1	1.5	1	1.5
2	3.5	3	5
2	3.5	3	5
3	5.5	1	1.5
3	5.5	4	7.5
4	7.5	3	5
4	7.5	2	3

We can extract the ranks, and calculate the Pearson coefficient for the ranks, getting $R_s = -0.10$ excluding the outlier.

x rank	y rank
1.5	7.5
1.5	1.5
3.5	5
3.5	5
5.5	1.5
5.5	7.5
7.5	5
7.5	3

Here's the data with the single outlier included. Again, I've assigned the rank to each value, with ties given the average rank.

x value	x rank	y value	y rank
1	1.5	4	7.5
1	1.5	1	1.5
2	3.5	3	5
2	3.5	3	5
3	5.5	1	1.5
3	5.5	4	7.5
4	7.5	3	5
4	7.5	2	3
10	9	10	9

We can extract the ranks, and calculate the Pearson coefficient for the ranks, getting $R_s = 0.24$ with the outlier included.

x rank	y rank
1.5	7.5
1.5	1.5
3.5	5
3.5	5
5.5	1.5
5.5	7.5
7.5	5
7.5	3
9	9

The Spearman coefficient is much less affected by the single influential point than is the Pearson correlation coefficient.