CHAPTER 3

# Correlation

Often economists are interested in investigating the nature of the relationship between different variables, such as the education level of workers and their wages or interest rates and inflation. Correlation is an important way of numerically quantifying the relationship between two variables. A related concept, introduced in future chapters, is regression, which is essentially an extension of correlation to cases of three or more variables that introduces an aspect of causality. As you will quickly find as you read through this chapter and those that follow, it is no exaggeration to say that correlation and regression are the most important unifying concepts of this book.

In this chapter, we will first describe the theory behind correlation, and then work through a few examples designed to think intuitively about the concept in different ways.

## Understanding correlation

Let $X$ and $Y$ be two variables (e.g. population density and deforestation, respectively) and let us also suppose that we have data on $i = 1, .. , N$ different units (e.g. countries). The **correlation** between $X$ and $Y$ is denoted by the small letter, $r$, and its precise mathematical formula is given in Appendix 3.1. Of course, in practice, you will never actually have to use this formula directly. Any spreadsheet or econometrics software package will do it for you. In Excel, you can use the Tools/Data Analysis or Function Wizard© to calculate them. It is usually clear from the context to which variables $r$ refers. However, in some cases we will use subscripts to indicate that $r_{XY}$

is the correlation between variables $X$ and $Y$, $r_{XZ}$ the correlation between variables $X$ and $Z$, etc.

Once you have calculated the correlation between two variables you will obtain a number (e.g. $r = 0.55$). It is important that you know how to interpret this number. In this section, we will try to develop some intuition about correlation. First, however, let us briefly list some of the numerical properties of correlation.

### Properties of correlation

1. $r$ always lies between −1 and 1, which may be written as $-1 \leq r \leq 1$.
2. Positive values of $r$ indicate a positive correlation between $X$ and $Y$. Negative values indicate a negative correlation. $r = 0$ indicates that $X$ and $Y$ are uncorrelated.
3. Larger positive values of $r$ indicate stronger positive correlation. $r = 1$ indicates perfect positive correlation. Larger negative values[1] of $r$ indicate stronger negative correlation. $r = -1$ indicates perfect negative correlation.
4. The correlation between $Y$ and $X$ is the same as the correlation between $X$ and $Y$.
5. The correlation between any variable and itself (e.g. the correlation between $Y$ and $Y$) is 1.

## Understanding correlation through verbal reasoning

Statisticians use the word correlation in much the same way as the layperson does. The following continuation of the deforestation/population density example from Chapter 2 will serve to illustrate verbal ways of conceptualizing the concept of correlation.

### Example: The correlation between deforestation and population density

Let us suppose that we are interested in investigating the relationship between deforestation and population density. Remember that Excel file FOREST.XLS contains data on these variables (and others) for a cross-section of 70 tropical countries. Using Excel, we find that the correlation between deforestation ($Y$) and population density ($X$) is 0.66. Being greater than zero, this number allows us to make statements of the following form:

1. There is a positive relationship (or positive association) between deforestation and population density.

2. Countries with high population densities tend to have high deforestation rates. Countries with low population densities tend to have low deforestation rates. Note that we use the word "tend" here. A positive correlation does not mean that **every** country with a high population density necessarily has a high deforestation rate, but rather that this is the **general tendency**. It is possible that a few individual countries do not follow this pattern (see the discussion of outliers in Chapter 2).
3. Deforestation rates vary across countries as do population densities (the reason we call them "variables"). Some countries have high deforestation rates, others have low rates. This high/low cross-country variance in deforestation rates tends to "match up" with the high/low variance in population densities.

All that the preceding statements require is for $r$ to be positive. If $r$ were negative the opposite of these statements would hold. For instance, high values of $X$ would be associated with low values of $Y$, etc. It is somewhat more difficult to get an intuitive feel for the exact number of the correlation (e.g. how is the correlation 0.66 different from 0.26?). The $XY$-plots discussed below offer some help, but here we will briefly note an important point to which we shall return when we discuss regression:

4. The degree to which deforestation rates vary across countries can be measured numerically using the formula for the standard deviation discussed in Chapter 2. As mentioned in point 3 above, the fact that deforestation and population density are positively correlated means that their patterns of cross-country variability tend to match up. The correlation squared ($r^2$) measures the proportion of the cross-country variability in deforestation that matches up with, or is explained by, the variance in population density. In other words, correlation is a numerical measure of the degree to which patterns in $X$ and $Y$ correspond. In our population/deforestation example, since $0.66^2 = 0.44$, we can say that 44% of the cross-country variance in deforestation can be explained by the cross-country variance in population density.

**Exercise 3.1**

**(a)** Using the data in FOREST.XLS, calculate and interpret the mean, standard deviation, minimum and maximum of deforestation and population density.

**(b)** Verify that the correlation between these two variables is 0.66.

### Example: House prices in Windsor, Canada

The Excel file HPRICE.XLS contains data relating to $N = 546$ houses sold in Windsor, Canada in the summer of 1987. It contains the selling price (in Canadian dollars) along with many characteristics for each house. We will use this data set extensively in future chapters, but for now let us focus on just a few variables. In particular, let us assume that $Y$ = the sales price of the house and $X$ = the size of its lot in square feet, lot size being the area occupied by the house itself plus its garden or yard. The correlation between these two variables is $r_{XY} = 0.54$.

The following statements can be made about house prices in Windsor:

1. Houses with large lots tend to be worth more than those with small lots.
2. There is a positive relationship between lot size and sales price.
3. The variance in lot size accounts for 29% (i.e. $0.54^2 = 0.29$) of the variability in house prices.

Now let us add a third variable, $Z$ = number of bedrooms. Calculating the correlation between house prices and number of bedrooms, we obtain $r_{YZ} = 0.37$. This result says, as we would expect, that houses with more bedrooms tend to be worth more than houses with fewer bedrooms.

Similarly, we can calculate the correlation between number of bedrooms and lot size. This correlation turns out to be $r_{XZ} = 0.15$, and indicates that houses with larger lots also tend to have more bedrooms. However, this correlation is very small and quite unexpectedly, perhaps, suggests that the link between lot size and number of bedrooms is quite weak. In other words, you may have expected that houses on larger lots, being bigger, would have more bedrooms than houses on smaller lots. But the correlation indicates that there is only a weak tendency for this to occur.

The above example allows us to motivate briefly an issue of importance in econometrics, namely, that of **causality**. Indeed, economists are often interested in finding out whether one variable "causes" another. We will not provide a formal definition of causality here but instead will use the word in its everyday meaning. In this example, it is sensible to use the positive correlation between house price and lot size to reflect a causal relationship. That is, lot size is a variable that directly influences (or causes) house prices. However, house prices do not influence (or cause) lot size. In other words, the direction of causality flows from lot size to house prices, not the other way around.

Another way of thinking about these issues is to ask yourself what would happen if a homeowner were to purchase some adjacent land, and thereby increase the lot size of his/her house. This action would tend to increase the value of the house (i.e.

an increase in lot size would cause the price of the house to increase). However, if you reflect on the opposite question: "will increasing the price of the house cause lot size to increase?" you will see that the opposite causality does not hold (i.e. house price increases do not cause lot size increases). For instance, if house prices in Windsor were suddenly to rise for some reason (e.g. due to a boom in the economy) this would not mean that houses in Windsor suddenly got bigger lots.

The discussion in the previous paragraph could be repeated with "lot size" replaced by "number of bedrooms". That is, it is reasonable to assume that the positive correlation between $Y$ = house prices and $Z$ = number of bedrooms is due to $Z$'s influencing (or causing) $Y$, rather than the opposite. Note, however, that it is difficult to interpret the positive (but weak) correlation between $X$ = lot size and $Y$ = number of bedrooms as reflecting causality. That is, there is a tendency for houses with many bedrooms to occupy large lots, but this tendency does not imply that the former causes the latter.

One of the most important things in empirical work is knowing how to interpret your results. The house example illustrates this difficulty well. It is not enough just to report a number for a correlation (e.g. $r_{XY} = 0.54$). Interpretation is important too. Interpretation requires a good intuitive knowledge of what a correlation is in addition to a lot of common sense about the economic phenomenon under study. Given the importance of interpretation in empirical work, the following section will present several examples to show why variables are correlated and how common sense can guide us in interpreting them.

**Exercise 3.2**

**(a)** Using the data in HPRICE.XLS, calculate and interpret the mean, standard deviation, minimum and maximum of $Y$ = house price (labeled "sale price" in HPRICE.XLS), $X$ = lot size and $Z$ = number of bedrooms (labeled "#bedroom").

**(b)** Verify that the correlation between $X$ and $Y$ is the same as given in the example. Repeat for $X$ and $Z$ then for $Y$ and $Z$.

**(c)** Now add a new variable, $W$ = number of bathrooms (labeled "#bath"). Calculate the mean of $W$.

**(d)** Calculate and interpret the correlation between $W$ and $Y$. Discuss to what extent it can be said that $W$ causes $Y$.

**(e)** Repeat part (d) for $W$ and $X$ and then for $W$ and $Z$.

## Understanding why variables are correlated

In our deforestation/population density example, we discovered that deforestation and population density are indeed correlated positively, indicating a positive relationship between the two. But what exact form does this relationship take? As discussed above, we often like to think in terms of causality or influence, and it may indeed be the case that correlation and causality are closely related. For instance, the finding that population density and deforestation are correlated could mean that the former directly causes the latter. Similarly, the finding of a positive correlation between education levels and wages could be interpreted as meaning that more education does directly influence the wage one earns. However, as the following examples demonstrate, the interpretation that correlation implies causality is not always necessarily an accurate one.

### Example: Correlation does not necessarily imply causality

It is widely accepted that cigarette smoking causes lung cancer. Let us assume that we have collected data from many people on (a) the number of cigarettes each person smokes per week ($X$) and (b) on whether they have ever had or now have lung cancer ($Y$). Since smoking causes cancer we would undoubtedly find $r_{XY} > 0$; that is, that people who smoked tend to have higher rates of lung cancer than non-smokers. Here the positive correlation between $X$ and $Y$ indicates direct causality.

Now suppose that we also have data on the same people, measuring the amount of alcohol they drink in a typical week. Let us call this variable $Z$. In practice, it is the case that heavy drinkers also tend to smoke and, hence, $r_{XZ} > 0$. This correlation does not mean that cigarette smoking also causes people to drink. Rather it probably reflects some underlying social attitudes. It may reflect the fact, in other words, that people who smoke do not worry about their nutrition, or that their social lives revolve around the pub, where drinking and smoking often go hand in hand. In either case, the positive correlation between smoking and drinking probably reflects some underlying cause (e.g. social attitude), which in turn causes both. Thus, a correlation between two variables does not necessarily mean that one causes the other. It may be the case that an underlying third variable is responsible.

Now consider the correlation between lung cancer and heavy drinking. Since people who smoke tend to get lung cancer more, and people who smoke also tend to drink more, it is not unreasonable to expect that lung cancer rates will be higher among heavy drinkers (i.e. $r_{YZ} > 0$). Note that this positive correlation does not imply that alcohol consumption causes lung cancer. Rather, it is cigarette smoking that causes cancer, but smoking and drinking are related to some underlying social attitude. This example serves to indicate the kind of complicated patterns of causality which occur in practice, and how care must be taken when trying to relate the concepts of correlation and causality.

### Example: Direct versus indirect causality

Another important distinction is that between **direct** (or immediate) and **indirect** (or proximate) causality. Recall that in our deforestation/population density example, population density ($X$) and deforestation ($Y$) were found to be positively correlated (i.e. $r_{XY} > 0$). One reason for this positive correlation is that high population pressures in rural areas cause farmers to cut down forests to clear new land in order to grow food. It is this latter on-going process of agricultural expansion which directly causes deforestation. If we calculated the correlation between deforestation and agricultural expansion ($Z$), we would find $r_{YZ} > 0$. In this case population density would be an indirect cause, and agricultural expansion, a direct cause of deforestation. In other words, we can say that $X$ (population pressures) causes $Z$ (agricultural expansion), which in turn causes $Y$ (deforestation). Such a pattern of causality is consistent with $r_{XY} > 0$ and $r_{ZY} > 0$.

In our house price example, however, it is likely that the positive correlations we observed reflect direct causality. For instance, having a larger lot is considered by most people to be a good thing in and of itself, so that increasing the lot size should directly increase the value of a house. There is no other intervening variable here, and hence we say that the causality is direct.[2]

The general message that should be taken from these examples is that correlations can be very suggestive, but cannot on their own establish causality. In the smoking/cancer example above, the finding of a positive correlation between smoking and lung cancer, in conjunction with medical evidence on the manner in which substances in cigarettes trigger changes in the human body, have convinced most people that smoking causes cancer. In the house price example, common sense tells us that the variable number of bedrooms directly influences house prices. In economics, the concept of correlation can be used in conjunction with common sense or a convincing economic theory to establish causality.

### Exercise 3.3

People with university education tend to hold higher paying jobs than those with fewer educational qualifications. This could be due to the fact that a university education provides important skills that employers value highly. Alternatively, it could be the case that smart people tend to go to university and that employers want to hire these smart people (i.e. a university degree is of no interest in and of itself to employers).

Suppose you have data on $Y$ = income, $X$ = number of years of schooling and $Z$ = the results of an intelligence test[3] of many people, and that you have calculated $r_{XY}$, $r_{XZ}$ and $r_{YZ}$. In practice, what signs would you expect these correlations to have? Assuming the correlations do have the signs you expect, can you tell which of the two stories in the paragraph above is correct?

## Understanding correlation through *XY*-plots

Intuition about the meaning of correlations can also be obtained from the *XY*-plots described in Chapter 2. Recall that in this chapter we discussed positive and negative relationships based on whether the *XY*-plots exhibited a general upward or downward slope.[4] If two variables are correlated, then an *XY*-plot of one against the other will also exhibit such patterns. For instance, the *XY*-plot of population density against deforestation exhibits an upward sloping pattern (see Figure 2.3). This plot implies that these two variables should be positively correlated, and we find that this is indeed the case from the correlation, $r = 0.66$. The important point here is that positive correlation is associated with upward sloping patterns in the XY-plot and negative correlation is associated with downward sloping patterns. All the intuition we developed about *XY*-plots in the previous chapter can now be used to develop intuition about correlation.

Figure 3.1 uses the Windsor house price data set (HPRICE.XLS) to produce an *XY*-plot of $X$ = lot size against $Y$ = house price. Recall that that the correlation between these two variables was calculated as $r_{XY} = 0.54$, which is a positive number. This pos-

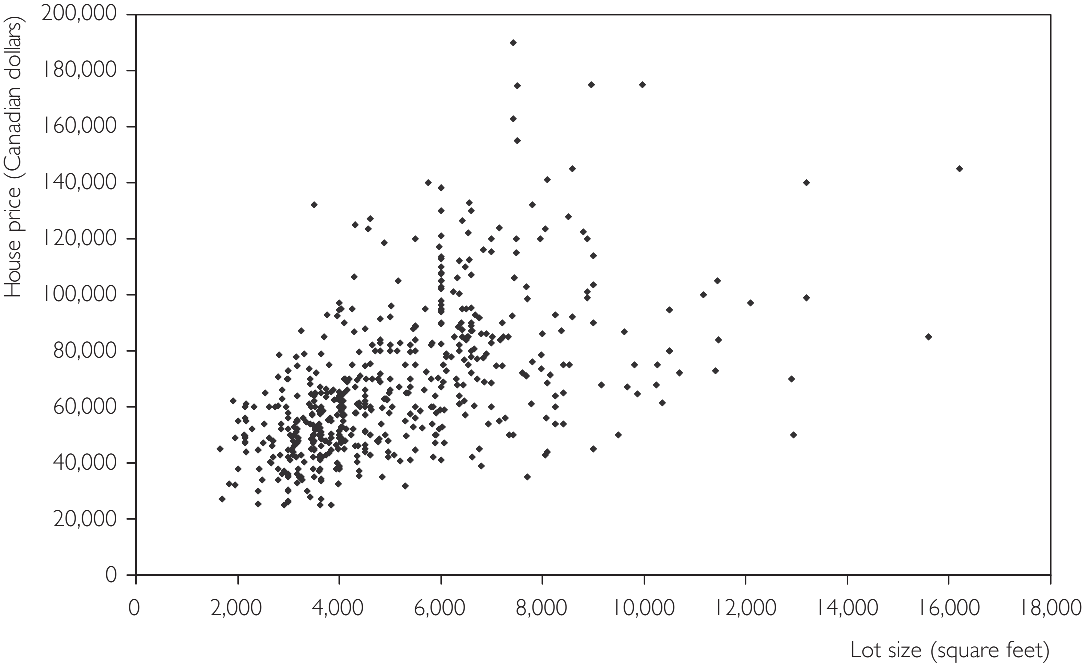

**Fig. 3.1** *XY*-plot of price versus lot size.

itive (upward sloping) relationship between lot size and house price can clearly be seen in Figure 3.1. That is, houses with small lots (i.e. small *X*-axis values) also tend to have small prices (i.e. small *Y*-axis values). Conversely, houses with large lots tend to have high prices.

The previous discussion relates mainly to the sign of the correlation. However, *XY*-plots can also be used to develop intuition about how to interpret the magnitude of a correlation, as the following examples illustrate.

Figure 3.2 is an *XY*-plot of two perfectly correlated variables (i.e. $r = 1$). Note that they do not correspond to any actual economic data, but were simulated on the computer. All the points lie exactly on a straight line.

Figure 3.3 is an *XY*-plot of two variables which are positively correlated ($r = 0.51$), but not perfectly correlated. Note that the *XY*-plot still exhibits an upward sloping pattern, but that the points are much more widely scattered.

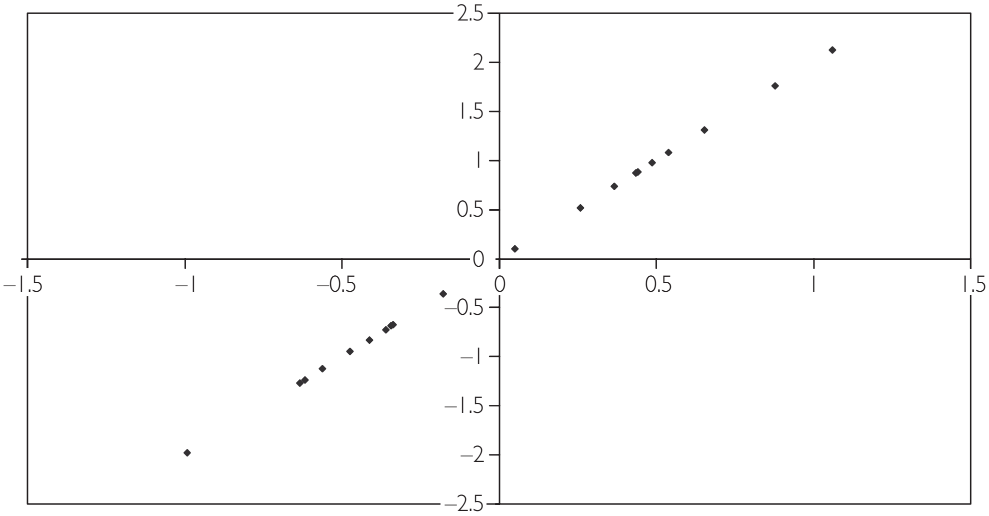

**Fig. 3.2** *XY*-plot of two perfectly correlated variables ($r = 1$).

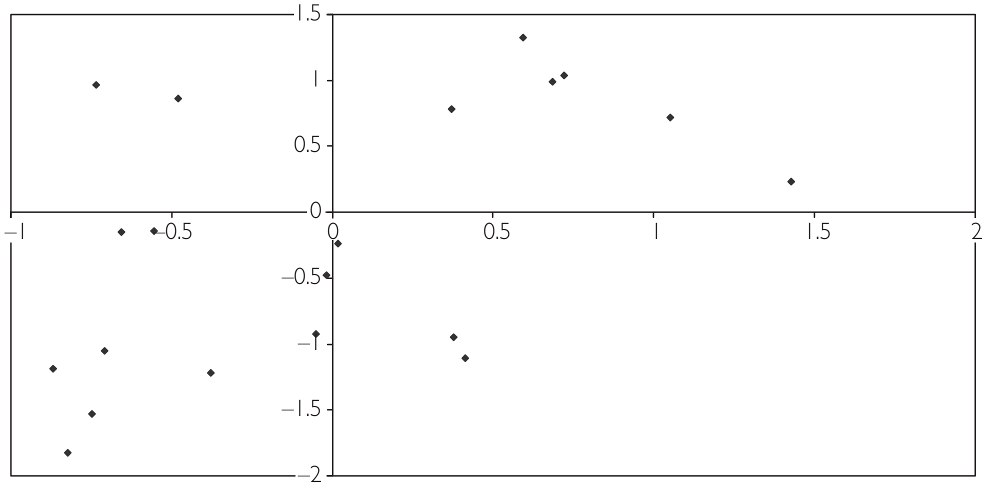

**Fig. 3.3** *XY*-plot of two positively correlated variables ($r = 0.51$).

Figure 3.4 is an $XY$-plot of two completely uncorrelated variables ($r = 0$). Note that the points seem to be randomly scattered over the entire plot.

Plots for negative correlation exhibit downward sloping patterns, but otherwise the same sorts of patterns noted above hold for them. For instance, Figure 3.5 is an $XY$-plot of two variables that are negatively correlated ($r = -0.58$).

These figures illustrate one way of thinking about correlation: correlation indicates how well a straight line can be fit through an $XY$-plot. Variables that are strongly correlated fit on or close to a straight line. Variables that are weakly correlated are more scattered in an $XY$-plot.

**Exercise 3.4**

The file EX34.XLS contains four variables: $Y$, $X_1$, $X_2$ and $X_3$.

**(a)** Calculate the correlation between $Y$ and $X_1$. Repeat for $Y$ and $X_2$ and for $Y$ and $X_3$.

**(b)** Create an $XY$-plot involving $Y$ and $X_1$. Repeat for $Y$ and $X_2$ and for $Y$ and $X_3$.

**(c)** Interpret your results for a) and b).

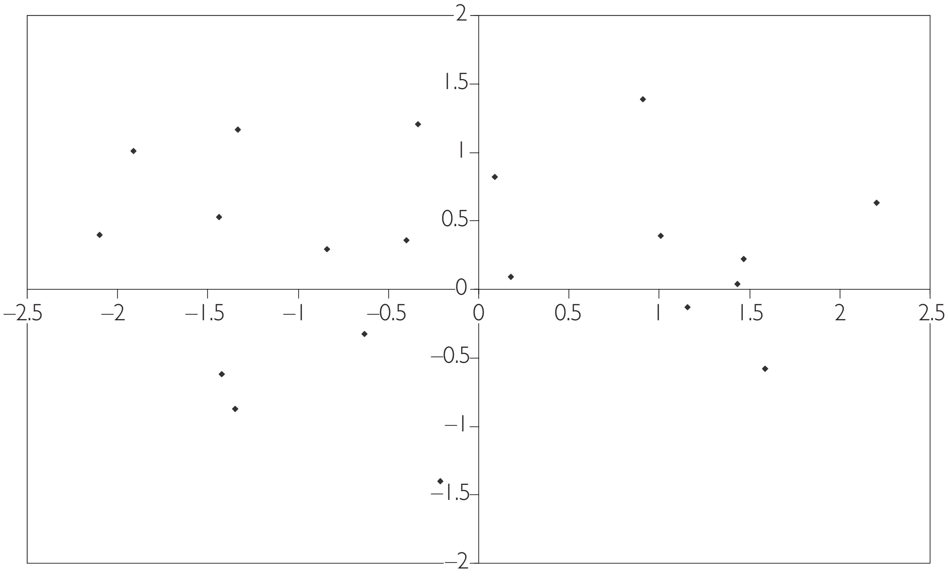

**Fig. 3.4** $XY$-plot of two uncorrelated variables ($r = 0$).

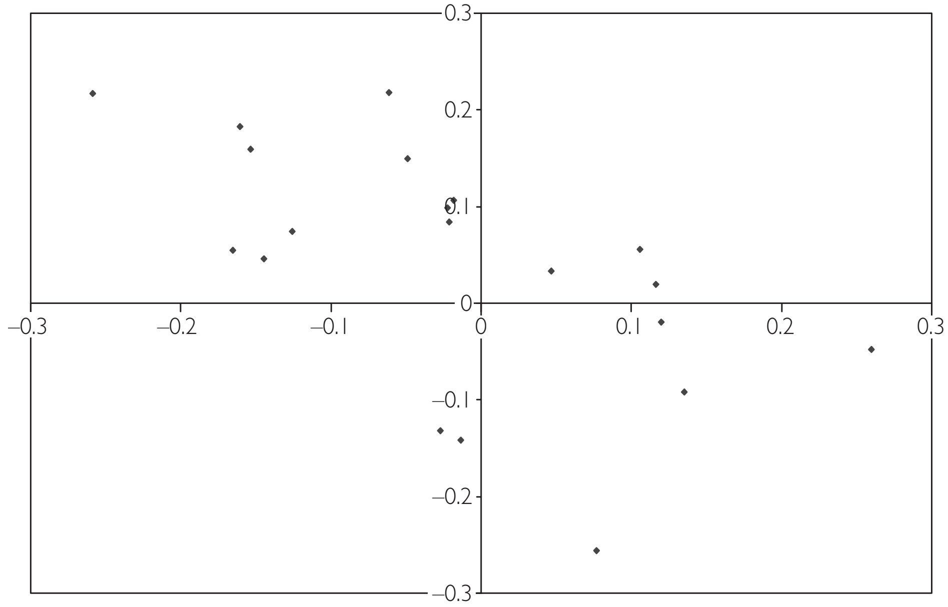

**Fig. 3.5** *XY*-plot of two negatively correlated variables ($r = -0.58$).

## Correlation between several variables

Correlation is a property that relates two variables together. Frequently, however, economists must work with several variables. For instance, house prices depend on the lot size, number of bedrooms, number of bathrooms and many other characteristics of the house. As we shall see in subsequent chapters, regression is the most appropriate tool for use if the analysis contains more than two variables. Yet it is also not unusual for empirical researchers, when working with several variables, to calculate the correlation between each pair. This calculation is laborious when the number of variables is large. For instance, if we have three variables, $X$, $Y$ and $Z$, then there are three possible correlations (i.e. $r_{XY}$, $r_{XZ}$ and $r_{YZ}$). However, if we add a fourth variable, $W$, the number increases to six (i.e. $r_{XY}$, $r_{XZ}$, $r_{XW}$, $r_{YZ}$, $r_{YW}$ and $r_{ZW}$). In general, for $M$ different variables there will be $M \times (M - 1)/2$ possible correlations. A convenient way of ordering all these correlations is to construct a matrix or table, as illustrated by the following example.

CORMAT.XLS contains data on three variables labeled $X$, $Y$ and $Z$. $X$ is in the first column, $Y$ the second and $Z$ the third. Using Excel, we can create a correlation matrix (Table 3.1) for these variables.

The number 0.318237 is the correlation between the variable in the first column ($X$), and that in the second column ($Y$). Similarly, –0.13097 is the correlation between $X$ and $Z$, and 0.096996, the correlation between $Y$ and $Z$. Note that the 1s in the correlation matrix indicate that any variable is perfectly correlated with itself.

**Table 3.1** The correlation matrix for X, Y and Z.

| | Column 1 | Column 2 | Column 3 |
|---|---|---|---|
| Column 1 | 1 | | |
| Column 2 | 0.318237 | 1 | |
| Column 3 | –0.13097 | 0.096996 | 1 |

**Exercise 3.5**

**(a)** Using the data in FOREST.XLS, calculate and interpret a correlation matrix involving deforestation, population density, change in pasture and change in cropland.

**(b)** Repeat part (a) using the following variables in the data set HPRICE.XLS: house price, lot size, number of bedrooms, number of bathrooms and number of storeys. How many individual correlations have you calculated?

## Chapter summary

1. Correlation is a common way of measuring the relationship between two variables. It is a number that can be calculated using Excel or any spreadsheet or econometric software package.
2. Correlation can be interpreted in a common sense way as a numerical measure of a relationship or association between two variables.
3. Correlation can also be interpreted graphically by means of *XY*-plots. That is, the sign of the correlation relates to the slope of a best fitting line through an *XY*-plot. The magnitude of the correlation relates to how scattered the data points are around the best fitting line.
4. Correlations can arise for many reasons. However, correlation does not necessarily imply causality between two variables.

## Appendix 3.1: Mathematical details

The **correlation** between *X* and *Y* is referred to by the small letter *r* and is calculated as:

$$r = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})(X_i - \overline{X})}{\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}},$$

where $\bar{X}$ and $\bar{Y}$ are the means of $X$ and $Y$ (see Chapter 2). More intuitively, note that if we were to divide the numerator and denominator of the previous expression by $N - 1$, then the denominator would contain the product of the standard deviations of $X$ and $Y$, and the numerator, the covariance between $X$ and $Y$. Covariance is a concept that we have not defined here, but you may come across it in the future, particularly if you are interested in developing a deeper understanding of the statistical theory underlying correlation.

# Endnotes

1. By "larger negative values" we mean more negative. For instance, −0.9 is a larger negative value than −0.2.
2. An alternative explanation is that good neighborhoods tend to have houses with large lots. People are willing to pay extra to live in a good neighborhood. Thus, it is possible that houses with large lots tend also to have higher sales prices, not because people want large lots, but because they want to live in good neighborhoods. In other words, "lot size" may be acting as a proxy for the "good neighborhood" effect. We will discuss such issues in more detail in later chapters on regression. You should merely note here that the interpretation of correlations can be quite complicated and a given correlation pattern may be consistent with several alternative stories.
3. It is a controversial issue among psychologists and educators as to whether intelligence tests really are meaningful measures of intelligence. For the purposes of answering this question, avoid this controversy and assume that they are indeed an accurate reflection of intelligence.
4. We will formalize the meaning of "upward" or "downward" sloping patterns in the *XY*-plots when we come to regression. To aid in interpretation, think of drawing a straight line through the points in the *XY*-plot that best captures the pattern in the data (i.e. is the best fitting line). The upward or downward slope discussed here refers to the slope of this line.