# APPROACHES AND METHODOLOGIES IN THE SOCIAL SCIENCES

## A PLURALIST PERSPECTIVE

EDITED BY

DONATELLA DELLA PORTA AND MICHAEL KEATING

CAMBRIDGE www.cambridge.org/9780521883221

# 13 Quantitative analysis

Mark Franklin

Quantification is one way of employing the scientific method to discover things about the world. In the social sciences we are trying to discover things about the social world, but the approach we use can still be regarded as scientific. The scientific approach attempts to abstract from the nuances and details of a story the salient features that can be built up into a theoretical statement (or statements) expected to hold true of any situation that can be defined in terms of the same abstractions. If such a theoretical statement does not hold true in some specific situation, this is presumed to be either because the theory was wrong or because it was not sufficiently elaborated. Elaborating social theories to bring in additional features of the world, found necessary for a full explanation, is an important feature of the scientific approach; but for elaboration to progress very far we need to employ quantitative analysis, as this chapter will try to show.

The transition from case studies to quantitative analysis is largely a matter of the number of cases. If you have one case, no causal inferences can be made. If you have two cases, you can rule out something as a necessary condition for something else. If you have three cases you can rule out two things, or you can start to make quantitative statements (for example, something might be found to pertain two-thirds of the time). As soon as you start saying things like 'this happens two-thirds of the time' you are doing quantitative analysis. But in order to make such statements you need to be able to abstract general features that are common to many cases, which tends to require a more elaborate theoretical basis for a quantitative study than for a case study. You also need a fairly large number of cases.

Exactly what constitutes 'fairly large' in the above statement is not at all clear, and in practice there is a large area of overlap in which one researcher would talk of a 'multiple case study' while another would talk of a 'small-*N* study' (the letter *N* in the quantitative tradition stands for 'number of cases'; as soon as you see cases referred to in that way, you know you are reading something written in the quantitative tradition).

**Table 13.1.** Governance and social networks

| | Multiple social networks going back to C12 | Lack of social networks even today |
|---|---|---|
| High-quality democratic governance | Northern Italy | |
| Poor governance | | Southern Italy |

*Source:* Adapted from Putnam (1993).

**Table 13.2.** Governance and social networks (after additional studies)

| | Multiple social networks going back to C12 | Lack of social networks even today |
|---|---|---|
| High-quality democratic governance | 2 | 0 |
| Poor governance | 1 | 3 |

So whether you do case studies or quantitative studies depends, over a large area of overlap, on what tradition you are working in rather than on what you are doing. Consider an example from Robert Putnam's (1993) study of democracy in Italy (Table 13.1 ). This rather famous example[1] was called into question only a few years later by a Harvard PhD thesis that looked at French regions and found a case of poor governance even where there were long-standing social networks. What to do? One possibility would be to conduct additional studies in the hope of discovering that either the French or the Italian findings were happenstansical – so unusual as to be not worth worrying about. One might, after a lot of work, come up with Table 13.2, where two cases of high-quality governance and three cases of poor governance match Putnam's findings, while the exception found in the just-mentioned thesis earlier turns out to be the only one.

That seems pretty definitive: Putnam's findings hold true far more often than not. Moreover, we can express the findings in terms of a condition that appears to be necessary for good governance (no examples of high-quality governance in Table 13.2 occur without it), even if that condition is not sufficient to ensure good governance.

It would, of course, be far more interesting to discover *why* the exception occurred, which would mean using the additional cases to see whether some other condition accounted for the exception. If we could find a magic ingredient (call it entrepreneurship) that accounted for the difference, we could make Table 13.3. This more elaborate test lets us see that there are actually two conditions, both of which must be present for high-quality democratic

**Table 13.3.** Entrepreneurship and networks

| | Entrepreneurship | | No entrepreneurship | |
|---|---|---|---|---|
| | Networks | Lack of networks | Networks | Lack of networks |
| High-quality governance | 2 | 0 | 0 | 0 |
| Poor governance | 0 | 1 | 1 | 2 |

**Table 13.4.** Territorial policy communities

| | Multiple social networks | | | |
|---|---|---|---|---|
| | Yes | | No | |
| Political entrepreneurship | Brittany | ✓ | Languedoc | × |
| | Tuscany | ✓ | | |
| Lack of entrepreneurship | Aquitaine | × | Provence | × |
| | | | Liguria | × |

*Source:* Smyrl 1997.

governance: the one found necessary by Putnam, and an additional condition he knew nothing about, which appears to be responsible for the exceptional French case. The additional condition turns out to be a second necessary condition for high-quality governance; entrepreneurship without multiple networks does not yield high-quality governance any more than do multiple networks without entrepreneurship.

Let me put the actual names of the regions concerned into a simpler table where the two conditioning variables determine where each region appears in the table, and the quality of governance in each region is indicated by a tick or a cross (Table 13.4 ). Even though both tables let us use the same logic of inference, Table 13.4 is the sort one would expect to see in a multiple case study, whereas Table 13.3 is the sort one would expect to see in a quantitative analysis. (In Table 13.4 I use the terminology of the author of the thesis.)

Of course, with only six cases, it is hard to be sure that one has exhausted the possibilities. Additional exceptions may lurk around the next corner, and additional conditions might need to be taken into account. But it is pretty obvious that to discover more one would need a great many additional cases, and with a great many additional cases the format used in Table 13.3 becomes more useful than that in Table 13.4. If we had dozens of names in Table 13.4 instead of only six, the information would not be very useful if presented in that format. With more than about ten cases, it becomes helpful to use

numbers to summarize what you have learned, trading off specificity for generality. But with small-*N* studies, what you can say with numbers is still quite limited. From this perspective, the next important watershed comes with the transition to 'large-*N* studies', where you can bring to bear the full power of what is called 'multivariate analysis'. Again, there is no fixed boundary. Small-*N* studies shade into large-*N* studies at somewhere between 30 and 300 cases, with progressively more powerful analyses being possible as *N* increases.

So what can be done with small-*N* studies that cannot be done with case studies, and what can be done with large-*N* studies but not with small-*N* studies? Essentially we can say that, as the number of cases goes up, so the researcher is better able to:

(a) specify the conditions under which causal effects are felt (how widespread they are);
(b) specify the nature of the causal effects (how strong they are);
(c) specify how likely it is that the effects are real rather than happenstansical (how significant they are).

## The vocabulary of quantitative research

The distinctions I have just made (among widespread, strong and significant causal effects) brings us to the main difficulty involved in quantitative analysis. To be able to talk quantitatively, one has to be able to make distinctions that to most people do not come naturally. Many of these distinctions, and the words used to make them, sound rather arbitrary. In ordinary English, the distinctions among strong, widespread and significant are not obvious. All appear to be variants on the word 'important'. That is true, but, as with the (perhaps apocryphal) fifty different words that the Inuit have for 'snow', distinctions that appear unimportant from some points of view can seem very important from other points of view.

In brief, a presumed causal effect is strong if it appears to have extensive effects. It is widespread if it occurs in many different circumstances and situations, and it is significant if it is unlikely to be spurious or happenstansical. When talking about accidents we use much the same vocabulary, distinguishing between a freak accident that probably will never happen again and one that is significant because it is part of a predictable pattern. But even a significant accident might have small or restricted consequences. Alternatively, its consequences could be major and/or widespread.

There is quite a lot of vocabulary to be learned in order to be able to talk

sensibly about quantitative social research findings, or to make sense of the literature that uses this vocabulary. In the rest of this chapter, I will go through some of the more important words concerned. Clearly, learning to do quantitative social research involves somewhat more than just learning the vocabulary. There are some corresponding skills, but I have always found that the vocabulary confuses people, rather than the skills. You may find it helpful to take a sheet of paper and write down the words in quotation marks that follow, to have a crib sheet to use as you move forward.

## Sources of quantitative information

Quantitative information can be collected in exactly the same way as any other information: by means of interviews (in the quantitative tradition these are generally called 'surveys') or by looking it up in compendia of various kinds (or on the Internet). Although there is no logical reason why this should always be true, surveys generally involve 'sampling' (we select a subgroup to interview because there are too many individuals for us to interview them all), whereas information that we look up is generally exhaustive (we can obtain data for the whole 'universe' of cases that interest us). It is important to know whether information was gathered from a sample rather than from a universe, because samples are subject to error when we try to generalize beyond the sample. This, of course, is equally true for many case studies, where the possibility of an 'unrepresentative case' is synonymous with a 'bad sample'; but there are certain types of sample ('probability samples') for which it is possible to use statistical methods to generalize beyond the sample with a known probability that the generalization will be true. This is a very powerful feature of 'random samples' that is unavailable to those who select their cases in other ways; in the case study tradition it is, strictly speaking, impossible to say how indicative a case might be. Most surveys are based on random sampling. Although there are different types of random sample, which need to be distinguished in practice, such distinctions are beyond the scope of this introduction.

## The dataset and data matrix

As soon as one starts talking about quantitative information, one is forced to start talking about data. Data (the word is plural – treating data as a collective noun is common but wrong) arise from standardized information. In this

sense, a biographical compendium contains data, because the characteristics of each individual are presented in a standard form: gender, birthdate, schools attended, and so on. A dataset goes further in coding the standardized data, generally in numerical terms (e.g. 1=female, 2=male) and providing a dictionary or 'codebook' with which to interpret the codes. When organized in this way, the codebook is conceptually distinct from the 'data matrix', which is a table organized with different cases in different rows. Across the table are columns, each column containing a particular characteristic (such as gender, age, income, or party voted for). These are known as 'variables'. By looking at the intersection of a particular row with a particular column, one can read off the particular characteristic or 'value' associated with a particular case. Thus if turnout at a European Parliament election were to be the variable in the third column of the table (Table 13.5) and France were the case in the fourth row, then by looking across the fourth row to the third column one would find that French turnout was 60.7 per cent at that election.

## Variables and levels of measurement

Talking about variables is complicated by the fact that there are different types of variable. Implicitly we have already mentioned two types: variables like gender, where the values ascribed are quite arbitrary, and variables like age, where the values ascribed have an intrinsic meaning (age is generally measured in years). In the case of a 'nominal' variable like gender, men could as easily be coded '1' and women '2' as the other way around – or the two possible values could be coded 'M' and 'F'. All we are doing with a nominal variable is distinguishing the characteristics that can apply to different cases in terms of that variable – the values we employ do no more than name the characteristics (hence 'nominal' from the Latin for 'name'). But with 'interval' variables like age, the intervals between the values are meaningful (a year or a dollar, or some other 'unit of measurement').

Two more levels of measurement are important to social science researchers. Variables can be 'ordinal' if the values have an order that is implied by their numeric values (5 is bigger than 4) even if there is no unit of measurement; and they can be 'dummy variables' if all they do is indicate the presence or absence of some characteristic (for example, 0=not British, 1=British). When the data come from a survey of individual people, the most common variables are nominal and ordinal, whereas the variables we really want in order to be able to conduct multivariate analyses (see below) are interval. A lot of time and

**Table 13.5.** European Election turnout (1)

Data matrix

| Country | Electn | EPturnout | Natturnout | Yrsleft | Compuls | First |
|---|---|---|---|---|---|---|
| bri | 1979 | 32.2 | 76 | 4 | 0 | 1 |
| den | 1979 | 47.8 | 86 | 0.36 | 0 | 1 |
| bel | 1979 | 90.4 | 95 | 2.4 | 1 | 1 |
| fra | 1979 | 60.7 | 83 | 2 | 0 | 1 |
| ger | 1979 | 65.7 | 91 | 1.3 | 0 | 1 |
| gre | 1981 | 82.2 | 82 | 0 | 1 | 1 |
| ire | 1979 | 63.6 | 76 | 2 | 0 | 1 |
| ita | 1979 | 84.9 | 91 | 4 | 1 | 1 |
| lux | 1979 | 88.9 | 89 | 0 | 1 | 1 |
| net | 1979 | 57.8 | 88 | 2 | 0 | 1 |
| bri | 1984 | 32.6 | 73 | 3 | 0 | 0 |
| den | 1984 | 52.4 | 88 | 3.2 | 0 | 0 |
| bel | 1984 | 92.2 | 95 | 1.3 | 1 | 0 |
| fra | 1984 | 56.7 | 71 | 1.7 | 0 | 0 |
| ger | 1984 | 56.8 | 89 | 2.6 | 0 | 0 |
| gre | 1984 | 82.2 | 82 | 0.96 | 1 | 0 |
| ire | 1984 | 47.6 | 73 | 2.7 | 0 | 0 |
| ita | 1984 | 83.4 | 89 | 3 | 1 | 0 |
| lux | 1984 | 88.8 | 89 | 0 | 1 | 0 |
| net | 1984 | 50.6 | 81 | 1.9 | 0 | 0 |
| por | 1984 | 72.4 | 73 | 0 | 0 | 1 |
| spa | 1984 | 68.9 | 70 | 2.4 | 0 | 1 |

Codebook

| Variables | Meaning (and values) |
|---|---|
| Country | Three-character country ID |
| Electn | Date of election (year)[1] |
| EPturnout | Turnout at European Parliament election (per cent) |
| Natturnout | Turnout at previous national election (per cent) |
| Yrsleft | Years to next national election (years and parts of years) |
| Compuls | Compulsory voting at time of EP election (0=no; 1=yes) |
| First | First EP election held in country (0=no; 1=yes) |

[1] Note that Greece, which held its first EP elections in 1981, is generally not distinguished from the 1979 election countries.

effort is expended by researchers in 'transforming' their data to overcome this problem. The solution generally adopted in political science research is to treat ordinal variables as interval (provided they have enough categories) and to recode nominal variables into their dummy counterparts, which can be regarded as honorary interval variables with a unit of measurement that is the

**Table 13.6.** Types of variable

| Level of measurement | Example | Additional information contained |
|---|---|---|
| Interval variable[1] | 43% Lab; 10% Lib Dem; 47% Con | Quantity (Con is 4% more than Lab) |
| Ordinal variable | 1=Lab, 2=Lib Dem, 3=Con | Order (left–right relative location) |
| Nominal variable | 1=Lab, 2=Con, 3=Lib Dem | Mutual exclusivity |
| Dummy variable | 0=Not Labour; 1=Labour | n.a. |

*Note:*

[1] Sometimes interval variables are further distinguished into those with a 'real zero point' which are called 'ratio scale' variables, but the distinction is not needed in the social sciences.

presence or absence of the attribute in question. This takes quite a lot of skill but, done properly, does not do violence to the data.[2]

In Table 13.6, not only do we see examples of different types of variables, but we also see a summary of the additional information needed to code a variable at a higher level than the level below it in the table; this is also the additional information imparted by such a coding. Dummy variables can be thought of as having the lowest level of information – the presence or absence of an attribute. Descriptions made in ordinary language generally consist of strings of attributes ('the man has blue eyes'). Talking of attributes enables us to string together different attributes of the same type ('the man has one blue eye and one green eye'). As soon as we move up to the nominal level, we assert that the attributes are mutually exclusive; one is allowed to vote only for a single political party, so a code of Conservative implies not Labour and not Liberal Democrat. By taking an additional step to the ordinal level, we introduce some additional concept that enables us to order the values – and also introduces the possibility of miscoding the variable according to this concept, as in the example in the table, where commentators argue about whether Labour and Liberal Democrat have recently swapped places in left–right terms.

Talking about variables requires us to make one further distinction, between variables we are trying to explain (dependent variables) and variables we are using in order to explain them (independent variables). In the example we used earlier, quality of democratic governance was the dependent variable because we were trying to answer the question 'What does the quality of governance depend on?'. Extensiveness of networks and the availability of entrepreneurial talent were independent variables because we were not (in that analysis) asking what they depended on. (Note that in some other piece of research one or other of those variables might very well be treated as dependent if, for example, we wanted to know what the availability of entrepreneurial talent depends on).

## Units and levels of analysis

Qualitative as well as quantitative analysis can focus on many different types of entity. One may analyse countries, years, regions, cities, schools, people or events – and much more. The entities we analyse are referred to as the units of analysis, or cases. The number of cases is referred to by the symbol *N*, as already mentioned. Units of analysis can be distinguished by the level of analysis at which they fall: the national unit is at a higher level than the city unit, which in turn is at a higher level than the individual who lives in that city and country. In Table 13.6, the example given of an interval variable is of a variable measured at a higher level, not only of measurement but also of analysis. To be able to say that Labour received 43 per cent of the vote, one has to be talking about an aggregation of individuals (most likely all of those voting at a particular election in a particular country) rather than of a particular individual. Because higher levels of analysis so often involve information about multiple individuals, the data concerned are often referred to as 'aggregate data'. The other examples in the table are ambiguous as to level of analysis (they could refer to political parties as easily as to individuals), but it is likely that they are variables measured at the individual level of analysis.

Although it is possible to investigate research questions that involve units at different levels of analysis, it is important to be clear about how these units are related to each other. This is just as true in qualitative as in quantitative studies, but with large-*N* studies it is easier to become confused about the level of analysis of different components of the study. The most important thing to realize about the level of analysis is that the types of variable we find at different levels tend to be different. I already mentioned that with individual-level data we get very few interval variables; in order to find an example of an interval-level variable relating to parties, I had to move up to an aggregate level of analysis. In addition, individual-level data generally contain a huge amount of error or 'noise'. People make mistakes when answering survey questions or when filling in forms. People fail to understand the questions they are asked or the meaning of the answers that they give. Most important, there is always a disjunction between the person who designs the questions (and hence the coding scheme for those questions) and the person who answers them (thus implicitly providing the values that will be coded). For this reason the questions often fail to communicate exactly the meaning intended. All of this results in error. There is generally much less error in higher-level data because individual-level error is averaged out during the process of aggregation. We

are also much more likely to find interval variables in aggregate data because the very act of aggregation yields variables that count the number (or proportion or percentage) of individuals in different categories or with different characteristics. The percentage voting Conservative (an aggregate phenomenon) is very definitely an interval variable, whereas the same variable at the individual level (voted Conservative) is a nominal variable, as we have already seen.

This might sound like a good reason to focus on aggregate rather than individual-level data, but there is a problem about deducing individual-level behaviour from aggregate-level data or vice versa. For instance, discovering that US states characterized by a high proportion of blacks in the population are states with a high proportion of illiteracy does not allow us to infer that blacks are more likely to be illiterate. In a famous article (Robinson 1950) it was found that in such states there was no difference between the literacy rates of whites and blacks. Both were less likely to be literate in states characterized by a high proportion of blacks. The error of inferring individual-level relationships from aggregate-level findings is called the 'ecological fallacy'. There is a corresponding 'individualistic fallacy' in inferring aggregate-level relationships from individual-level relationships. For example, the strong positive relationship found at the individual level between education and voting does not translate into a corresponding positive relationship at the national level. To the contrary, the two countries with among the best education systems on earth (the United States and Switzerland) have among the lowest rates of voter turnout (Franklin 2004).

So data need to be collected and analysed at the level of analysis appropriate to the research question that is being asked, and analysts should avoid making generalizations at a different level of analysis from the level of the data that gave rise to the findings. This requirement is an instance of a more general requirement, common to all types of investigation (quantitative or qualitative), of thinking carefully about how variables are measured and about the inferences that can be made from different types of variable used in different ways. Measurement error is always a threat to inference, whether in qualitative or quantitative work (see King, Keohane and Verba 1994).

## Statistics

In order to talk about quantitative research findings, one needs to use statistics. Technically speaking, statistics are 'coefficients' that summarize things of

interest about data. Statistics are also the procedures by which one arrives at such coefficients, generally referred to by those who do it as 'statistical analysis'. A percentage or an average is a statistical coefficient (generally referred to as a 'descriptive statistic' because it describes a body of data), but much more interesting to social scientists are coefficients that address the questions summarized earlier: How widespread? How strong? How significant? We will start with the last of these.

## How significant?

'Significance' relates to the chances of being wrong when making some assertion. Statistical methods allow us to determine the chances of being wrong about conclusions reached from a random sample. By extension, most researchers apply these methods to any dataset for which there is no reason to doubt its representative nature. Questions of significance can be applied to what are called 'point estimates' (for example, statistics can tell us how likely it is that we are wrong if we estimate that the Democrats will win 53 per cent of the two-party vote at the next US presidential election); but much more interesting to social scientists are questions about the significance of a relationship between variables. If we take the example, used earlier, of the relationship between the extent of policy networks and the quality of governance, it would be worth knowing the chances that the relationships found by Putnam and Smyrl are significant ones – that is, that they are unlikely to be the result of happenstance and are thus likely to be found again and again as we look at other regions and countries.

Whether a relationship is significant depends on three things:

(1) the strength of the relationship;
(2) the number of cases investigated when establishing the relationship;
(3) the degree of certitude required before we are willing to accept a statement as true.

Starting with the last of these, if we require 100 per cent certitude (generally referred to as 'confidence'), it will follow that no relationship is significant. Virtually all social science statements are probabilistic by nature (whether discovered using quantitative or qualitative methods). The industry standard in the quantitative social sciences is to accept a statement as true if it is likely to be correct in 95 per cent of the instances to which it might be generalized, which is the same as saying that the statement will be false in 5 per cent of these instances – for which reason it is referred to as 'significance at the 0.05 level'.

Note that this is not a very stringent test. If 5 per cent of situations to which a finding might be generalized will fail to show the relationship concerned, this means that one in twenty situations will fail to show it. Equally, if we cannot establish a finding at the 0.05 level of significance, then there is still a one in twenty chance that the relationship in question is nevertheless real. If we want greater certitude, we need to conduct a more stringent test; for instance, requiring significance at the 0.01 level, which would imply being wrong only once in a hundred times when generalizing from the finding. But for this we need more cases, as will now be explained.

If we want to be able to assert that there is a relationship between the extensiveness of social or policy networks and the quality of democratic governance, the more cases we have investigated in arriving at that assertion, the better. If we examined every single relevant case and found that all of them showed the same relationship, we would be pretty confident about our assertion. With a proper random sample of cases, we can say how confident we are that all the unexamined cases would show the same relationship as that found among the cases that were investigated. Enough cases can render any relationship significant at any non-zero level of significance, so with enough cases the question of significance ceases to be very interesting; but in general, the more cases the better.

However, it is also important to realize that, even with a relatively small *N*, relationships can prove significant if they are strong enough, which is the third thing needed for significance (the first one as listed above). As should already be clear, it takes many cases to establish that a weak relationship is significant, while a very strong relationship can be established even with relatively few cases. In the unusual situation where we expect definitive relationships of the kind 'all X's are Y's' or 'no X is ever a Y', we only need enough cases to rule out measurement error. If we expect to find a less deterministic relationship (and most relationships in the social sciences are probabilistic rather than deterministic, as mentioned), then we need more cases in order to be confident of our findings.

## How strong?

To determine how strong a relationship is, we must determine the amount of change in the dependent variable that is brought about by change(s) in the independent variable(s). A small change is much more likely to be happenstansical than a large change, but more importantly, a small change is not very interesting even if it were to prove significant. When talking about strengths

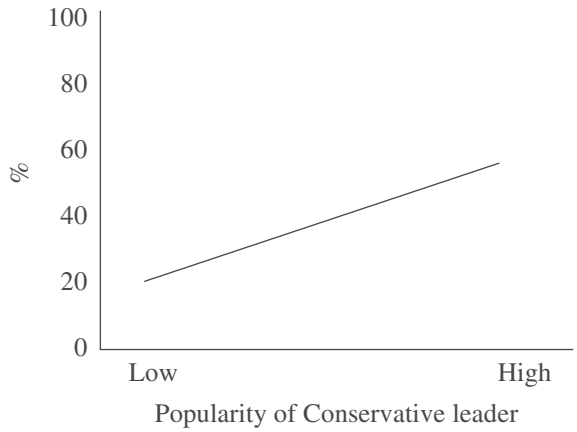


**Fig. 13.1** Chances of a Conservative electoral victory

of relationships it helps to think of a graph that has the dependent variable arrayed up and down the vertical axis and an independent variable arrayed along the horizontal axis. For a given movement across the horizontal axis we can then read off the corresponding movement up the vertical axis, as shown in Figure 13.1.

In that graph, we see the chances of a Conservative victory increasing from only 20% to about 60% as the popularity of the Conservative leader increases from low to high. This corresponds to a 40% difference (60% – 20%, or an 'effect' of 0.4, since effects are generally expressed as proportions). One can think of the slope of the line in terms of the leverage it shows the independent variable having on the dependent variable. An almost flat line corresponds to very little leverage. A strongly sloping line corresponds to much more leverage. An effect of 0.4 gives quite a lot of leverage. By contrast, it is clear that an effect of only 0.04 (4%) would yield a line that was almost flat – a line with almost no leverage. A downward slope is also possible and would indicate a negative relationship: increasing values of the independent variable would correspond to decreasing values of the dependent variable.

The relationship shown in a table (such as those we used earlier) can easily be converted to a graph such as the one in Figure 13.1 by percentaging the table in the direction of the dependent variable. Thus, in Table 13.2 above (the first of those relating to Putnam's theory that contained any numbers), the dependent variable (quality of governance) runs down; so we percentage down and find that 67% of regions with extensive social networks (2 out of 3) see high-quality governance, whereas 0% of regions without extensive networks

see high-quality governance. Subtracting, we find that social networks make a difference of 67 – 0 = 67% to the quality of governance (i.e. social networks have an effect on governance of 0.67). That is a pretty strong effect on a scale that goes from 0 to 1 which, if turned into a graph, would show a slope even steeper than the one depicted in Figure 13.1. The steepness of the slope in this example corresponds to our intuition that a single exception to Putnam's asserted rule does not amount to much; but the small number of cases would preclude even so strong an effect from being statistically significant even if the cases had been chosen randomly.

## Correlations between variables

At this point, we need to take a brief detour to talk about correlations. Rather than referring to the effect of one variable on another, when dealing with only two variables social scientists often talk about the 'correlation' between them, generally denoted by the symbol *r* (or sometimes *R*). *R* stands for 'relationship', and talking about relationships between variables does not require us to distinguish between dependent and independent variables. Two variables are related if their values tend to move together (taller people tend to be heavier so there is a relationship between height and weight). There is also said to be a relationship – a negative relationship – if two variables tend to move inversely (the thicker the clouds, the dimmer the daylight). If both variables are scaled between 0 and 1 (or, in general, both are measured on the same scale), then measures of correlation will take on approximately the same values as the effects we have been talking about. The effects of 0.4 and 0.67 to which we have referred would correspond to correlations of 0.4 and 0.67, or very nearly. Correlations are preferable for some purposes, however, because the value of a correlation coefficient does not depend on the scale of measurement of the variables concerned. If we were investigating the relationship between age in years and income in euros, the effect of age on income would certainly be far greater than 1.0 (a one-year increase in age would generally result in several hundred more euros in income) and would be hard to interpret, whereas the correlation coefficient would be somewhere in the range –1.0 to +1.0, just like the coefficients we have been discussing. Table 13.7 shows the approximate substantive meaning to be ascribed to correlation coefficients of different magnitudes when using individual-level and aggregate data (boundaries are not hard and fast and would be disputed by some analysts).

**Table 13.7.** Strength of correlation

| Strength of correlation | Interpretation with individual-level data | Interpretation with aggregate data |
|---|---|---|
| $r/R = 0.00–0.06$ | Trivial | Trivial |
| $r/R = 0.07–0.19$ | Slight | Trivial |
| $r/R = 0.20–0.34$ | Moderate | Slight |
| $r/R = 0.35–0.49$ | Strong | Moderate |
| $r/R = 0.50–0.65$ | Spectacular | Strong |
| $r/R = 0.66–0.80$ | Highly spectacular | Very strong |
| $r/R = 0.81–0.95$ | Suspect | Spectacular |
| $r/R = 0.96–1.00$ | Very suspect | Suspect |

*Note:* Interpretations apply to $r$ for bivariate analysis, $R^2$ for multivariate analysis (see below).

As stated earlier, it is difficult to find strong relationships using individual-level data because those who design the question categories are generally not those who answer the questions, so that any number of misunderstandings can result. Also, individuals are frequently quite uncertain about how to answer even questions that they correctly understand, and often cannot be bothered to think carefully about their answers. This results in a great deal of error that is largely absent from aggregate data, or is averaged out when individual-level information is aggregated. Thus, we expect stronger correlations (and stronger effects) with aggregate data than with individual-level data. Indeed, individual-level correlations above 0.8 are so unusual as to generally suggest that something about the analysis was done wrongly, or something about the data is not quite right. Very often in such cases the analyst has employed two variables that are in reality different measures of the same thing, so that the finding is tautological. With aggregate data, correlations above 0.9 are quite attainable (though unusual), and only correlations above about 0.95 suggest the testing of tautological relationships.

## How widespread?

The extent to which a relationship is widespread is a matter of the number of situations in which it is found. A relationship found only where there are extensive social networks is less widespread than one which is also found where social networks are absent. Establishing how widespread is a relationship requires the use of multiple independent variables in order to specify the different circumstances in which that relationship does or does not hold. In

the Putnam example we started with, the relationship between networks and governance held only in the case where entrepreneurship was present, so this relationship proved not to be as widespread as originally supposed by Putnam. A relationship that holds only in certain circumstances is said to be subject to an 'interaction'. In this case there was an interaction between entrepreneurship and the extent of social networks, such that each had its effect only in the presence of the other. In order to test for interaction effects it is necessary to employ a great many independent variables, one for each of the circumstances in which an effect might or might not be found to hold true. But we need multiple independent variables for another reason as well, to which we now turn.

## Multivariate analysis

So far, except when examining the Putnam thesis, we have been talking only about so-called 'bivariate' relationships: relationships that may be found when a single dependent variable is examined in relation to a single independent variable. It is unusual to be able to explain much about the world with bivariate relationships, partly because there is so much error in our data (especially in our individual-level data) – error that often needs to be measured and specified in order to correctly estimate the effects of the variables of interest.[3] More importantly, the social world is a complex place. All the circumstances that might affect how widespread a relationship is (as just explained) may also contribute to an explanation of the dependent variable of interest. When we bring additional independent variables to bear in this way we are said to be 'elaborating' our explanation, as mentioned in the opening paragraph of this chapter. Indeed, the need to take account of multiple simultaneous effects on a dependent variable occurs in practice more frequently than the need to take account of interaction effects. But as soon as we move beyond bivariate analysis we need new tools for thinking about relationships, and when we use such tools we are said to be performing 'multivariate analysis'.

Strictly speaking, the analysis we performed in Tables 13.3 and 13.4 were multivariate analyses because more than a single independent variable was involved. However, the tools we used (tables, percentages, percentage differences) were the tools of bivariate analysis. When we move to multivariate analysis proper we need to think of effects in terms of equations, and this is another step that many budding social scientists find quite daunting – unless it is explained to them that equations are perfectly straightforward tools that

everyone uses implicitly every time they add up the charges they expect to incur on their next mobile phone bill.

A typical mobile phone bill has a total that is the result of taking a standing monthly charge and adding to it an amount for calls in excess of some maximum, perhaps an amount for roaming, perhaps an amount for taxes, and so on. The result is a sum that can be spelled out as an equation such as:

$$\text{Total due} = \text{standing charge} + \text{minutes}^{*}\text{charge}_{\text{per minute}} + \text{roaming}^{*}\text{charge}_{\text{per roaming minute}}$$

(perhaps with another component for taxes). In the equation, the plus sign signifies addition and the asterisk signifies multiplication. People find it fairly straightforward to multiply the number of minutes by the charge per minute and the number of roaming minutes by the charge per roaming minute and add those two products to the standing charge. What gives them trouble is when the words used in the above equation are replaced with symbols, as in the following:

$$Y = a + b_1X_1 + b_2X_2$$

Here the total due is replaced by the symbol *Y*, the standing charge by the symbol *a*, the number of excess minutes by the symbol *X* with a subscript of 1, and the number of roaming minutes by the symbol *X* with a subscript of 2. Each *b* is the charge per minute for the corresponding number of minutes (again with the appropriate subscript).

The use of symbols in place of words looks quite cumbersome but is actually very powerful. By convention we always use the symbol *Y* to stand for the dependent variable and *X* (with different subscripts) to stand for different independent variables. Each *b* measures the effect of the relevant *X* on the dependent variable. The symbol *a* is always used to denote a constant, which might be zero if, in an example such as the telephone bill, there was no standing charge. Evidently we can extend the equation with many more *X*'s without running out of space on the line, and we can talk conceptually about what we are doing without having to use any specific examples of actual variables. In the Putnam example, we could write the equation that we were implicitly evaluating exactly as above, where *Y* stands for the quality of democratic governance, $X_1$ for the extent of social/political networks and $X_2$ for the availability of entrepreneurship. In practice, in this example the constant (*a*) term was implicitly zero because the quality of governance was so poor in the absence of the two necessary conditions.[4] Note that we cannot actually estimate the effects inherent in Table 13.2 without considerable gyrations. The

only effect we calculated – the 0.67 effect of networks on quality in the case where entrepreneurship was available – is what is called a 'partial effect', an effect that applies only in a specified circumstance.

In order to calculate effects of independent variables on dependent variables in a multivariate analysis, several methods are available; but the most widely used is called 'regression analysis'.

## Regression analysis

This type of analysis gets its name, in a most unlikely way, from the fact that it was developed by geneticists to study the way in which offspring who are taller or shorter than their parents tend themselves to have children whose height 'regresses towards the mean'. In this brief introduction there is no need for us to explain how the calculations are performed. All that is necessary is to know that, for any given dependent variable *Y*, regression analysis produces values for the constant *a*, and for each of the *b*'s used in investigating the relationships concerned. The analyst must supply the data for *Y* and for each of the *X*'s, which will generally be contained in a data matrix such as the one presented earlier. Using those same data, from Table 13.5, we can investigate whether the level of turnout at European Parliament elections for different countries is predictably related to turnout at each country's most recent national election together with the length of time until its next national elections, along with a correction for compulsory voting (countries with compulsory voting see much less drop-off in turnout at European Parliament elections than other countries do). The results can be expressed in this equation:

EPturnout = 24.7 + 0.30*Natturnout + 32.9*Compuls + 7.2*First

This equation would tell us that there is a floor to turnout at European Parliament elections of about 25%, to which can be added a small proportion (0.30) of the turnout at the previous national election, but with a correction that adds almost 33% in countries with compulsory voting, and another 7.2% in the case of the first European Parliament elections ever conducted in the country concerned.

Of course, extracting that information from the output of a statistical package is not totally straightforward. Table 13.8 reproduces a portion of that output from a typical software package – output giving rise to the equation above. The names of variables appear down the left-hand side (dependent variable at the top). The coefficients in the next column are those used in the

**Table 13.8.** European Election turnout (2)

| EPturnout | Coeff. | s. e. | *t* | Prob |
|---|---|---|---|---|
| Natturnout | 0.30 | 0.18 | 1.66 | 0.10 |
| Compuls | 32.90 | 3.30 | 9.95 | 0.00 |
| First | 7.16 | 2.90 | 2.47 | 0.02 |
| (Constant) | 24.67 | 14.02 | 1.76 | 0.08 |
| Number of observations | | | | 64 |
| F(3, 60) | | | | 66.38 |
| Prob > *F* | | | | 0.00 |
| $R^2$ | | | | 0.77 |
| Adjusted $R^2$ | | | | 0.76 |

equation. Other coefficients are described later or are beyond the remit of this chapter, but the column headed Prob (sometimes Prob is abbreviated to P) gives the level of significance of each effect. The fact that the effect of Natturnout has a probability of 0.10 of being spurious tells us that European Parliament turnout is probably not in fact affected by turnout at the previous national election, so that this component of the equation should in practice be eliminated (and will be eliminated in Table 13.9, as our story proceeds).

The output from the regression program also tells us the $R^2$ associated with the analysis, among many other statistics. The $R^2$, not surprisingly, is the square of $R$ (or $r$) – the coefficient often used to describe bivariate relationships that was discussed earlier. The value is squared in multivariate analysis partly because, with more independent variables, it is easier to achieve a high value of $R$. By squaring this coefficient, one arrives at a smaller coefficient more appropriate for use in multivariate analysis (a proportion of a proportion is a smaller proportion – for example, a half of a half is a quarter). To evaluate values of $R^2$, one can use Table 13.7 for interpreting different values of $r$. A spectacular individual-level multivariate finding is one that yields an $R^2$ above 0.5, whereas with aggregate data the $R^2$ would have to be above 0.8 to be spectacular, and so on. Table 13.8 also lists an adjusted $R^2$, which is the value generally reported.

In the remainder of this section, we will describe the analysis that followed from the discovery (illustrated in Table 13.8) that turnout at European Parliament (EP) elections was not significantly affected by turnout at the previous national election for each country. This finding came as quite a surprise, because EP elections are supposed to be secondary to *national* elections (Reif and Schmitt 1980), demonstrating features of the national situation rather than features pertaining to the EP election itself. Thus, although it is natural

**Table 13.9.** European Election turnout (3)

| | Model A | | Model B | | Model C | | Model D | |
|---|---|---|---|---|---|---|---|---|
| Independent variables | *b* | (s.e.) | *b* | (s.e.) | *b* | (s.e.) | *b* | (s.e.) |
| Natturnout | 0.30 | (0.18) | | | | | | |
| Compuls | 32.90 | (3.30)** | 36.22 | (2.66)** | 38.30 | (2.98)** | 38.62 | (2.74)** |
| First | 7.15 | (2.90)* | 8.30 | (3.86)* | 1.51 | (5.38) | | |
| First*NotCompuls | | | | | 9.41 | (6.34) | 10.92 | (3.31)** |
| (Constant) | 24.67 | (14.02) | 47.80 | (1.62)** | 47.15 | (1.66)** | 47.14 | (1.65)** |
| Adjusted $R^2$ | 0.76 | | 0.75 | | 0.75 | | 0.76 | |
| *N* | 64 | | 64 | | 64 | | 64 | |

*Note:* Dependent variable is EPturnout; $p$ = *0.05, **0.01.

to theorize that a primary determinant of EP election turnout is national election turnout, the relevant coefficient is not significant in Table 13.8.

Table 13.9 presents the findings of a series of different regression analyses (described as 'models' in the table), each one using slightly different independent variables, in order to step the reader through the findings that led to the rejection of the intuitively more appealing theory and the acceptance of a model (which might be quite surprising to some) that makes no use of national election turnout as an independent variable. The table is laid out in a fashion customary in contemporary journal articles, with the names of the independent variables down the left-hand column and then a pair of coefficients for each variable for each model. The first in each pair of coefficients for each model is the coefficient of primary interest – the *b* coefficient that might be taken from the output of a computer program (such as illustrated in Table 13.8) and transferred to an equation (such as the one presented earlier). The second coefficient in each pair is headed s.e. (which stands for 'standard error' – coefficients that can also be found in Table 13.8), which measures how much error there is in each *b* coefficient; sometimes the parenthesized standard error appears under its corresponding b coefficient. It is not important for the purposes of this chapter to understand these coefficients, but they are used to determine the level of significance of the effect (the Prob coefficients in Table 13.8), which in published tables that look like Table 13.9 are generally indicated by one or more stars following the coefficient. The critical question those coefficients answer is 'How much error is there in the *b* coefficient relative to its size?'; as the amount of error approaches or exceeds the size of the coefficient, so significance is reduced. In Table 13.9, coefficients are given one star to show that they are significant at

the 0.05 level and two stars to show that they are significant at the 0.01 level, but other conventions are also seen.[5] The meaning ascribed to the stars is always given in a footnote to the table. When the data come from a random sample, we stand only a 1 in 100 chance of being mistaken when we assert that effects with two stars are real. In the last two rows, at the foot of each model, are presented the number of cases included in the analysis ($N$) and the $R^2$ associated with the analysis, which we have already described in connection with Table 13.8.

Based on this rather minimal introduction, we can proceed to explain why the intuitively more appealing notion (that turnout at EP elections would depend on turnout at national elections) was rejected in favour of an explanation that does not even mention national elections. Model A is the model already presented in Table 13.8, repeated for reference purposes. This is the theoretically expected model in which, however, national turnout proves not significant (no stars for the effect of 0.30). In Model B, we see what happens when we simply remove national turnout from the model. The other variables increase their effects a little, but the effect of first election is still significant only at the 0.05 level, and the variance explained (adjusted $R^2$) goes down a bit. Some thought suggests that perhaps we are misspecifying our first election variable, because theoretically the fact that there is something special about an election should not affect turnout in a country that already has compulsory voting. Specifying an appropriate interaction between first election and compulsory voting, in addition to first election, yields a model (Model C) in which neither of these variables proves significant, but the interaction effect is by far the stronger of the two effects. Since first election was significant when it was the only measure of the concept (in Model B), its failure to prove significant when accompanied by its new variant (in Model C) must be because the two variables are largely measuring the same thing (this is called 'multicollinearity'). There are several ways to deal with multicollinearity, but in this example we address it by simply eliminating the less powerful of the two alternative measures. The result is model D, where all effects are highly significant and variance explained is back up to where it was in Model A.[6] (For a detailed presentation of these ideas, see Franklin 2001.)

## The way forward

There is much still to learn about quantitative analysis. In particular, there are a great many types of multivariate analysis, many of them designed

for specialized research situations, with the choice among them being largely dictated by the nature of the data being analysed. For example, data in which the cases constitute different points in time require a whole set of specialized procedures, as do data measured at different levels of aggregation.

Nevertheless, regression analysis is something of an 'industry standard' for multivariate analysis. Being able to understand the coefficients presented in published research papers that derive from regression analysis (together with the vocabulary used to describe those coefficients and the analyses that give rise to them) will take budding social scientists a long way. Being able to 'do' regression analysis in their own research will help them to be critical consumers of such research findings. Such relatively straightforward skills will also cover a large majority of the situations they are liable to encounter in the world of quantitative research.

This chapter has also illustrated a feature of quantitative analysis that is frequently overlooked. It is often stated that the scientific method proceeds deductively by testing propositions derived from theories that originate elsewhere (see Héritier, ch. 4). More typical of scientific research (not just in the social sciences) is, however, the example given in the previous section of how our understanding of turnout in European Parliament elections was elaborated. Scientists do not use data only to test their theories. They also use data to revise their theories and/or arrive at new ones. Archimedes discovered his Principle by observing his bathwater overflow, and virtually every scientific discovery is based ultimately on observation. Sometimes the observations concerned are direct (as with Archimedes or Putnam) and sometimes they are indirect, based on analysis of data collected for other purposes, as in the example reported in Table 13.9. This very important distinction is referred to elsewhere in this volume.

A huge part of what we know about the world is based on data analysis, and this is especially true in the social sciences. In these disciplines, relationships are often so complex that many variables need to be observed and manipulated simultaneously in order to control for all the things going on in the world that are not of primary interest but that could contaminate our findings. Often, a clear view can only be obtained by means of quantitative analysis of the data. That clear view will generally be at a high level of abstraction but, even though abstract, it can help greatly in the understanding of specific developments in particular places: it can help those conducting case studies to decide what to focus on,[7] just as much as case studies can help quantitative researchers decide what to measure.

## NOTES

1 Robert Putnam's *Making Democracy Work* established the concept of 'social capital' within the contemporary literature of political science (it originated in the work of sociologist James Coleman). Putnam himself developed it in his later book *Bowling Alone*, but the ideas in *Making Democracy Work* were also picked up by other political scientists so that social capital studies have become something of a growth industry in recent years.

2 The mutual exclusivity of nominal-level variables (see below) is not something we find very useful to know, so losing this information does not cost us much. Pretending there is a unit of measurement for an ordinal-level variable equally need cost us little in practice.

3 This can be thought of in terms of measuring the various contaminants that would otherwise threaten the reliability of quantitative findings. In some of the natural sciences, contamination can be ruled out by careful cleaning of scientific instruments. In the social sciences, contaminants must be measured and relevant indicators included in any analysis that hopes to arrive at correct (what econometricians call 'unbiased') results. Many of the variables included in multivariate analyses are of no interest on their own account but are included because they are known to affect the dependent variable, and to leave them out would result in 'omitted variable bias'. Measuring and including contamination can even substitute for the use of proper random samples if the sources of error are sufficiently extensively specified.

4 Actually, that might not be true. The need to specify a constant term in an equation draws attention to something missing from the common characterization of Putnam's findings. Presumably the quality of governance in southern Italy was not zero, and perhaps was different in different southern regions, pointing to the need to elaborate Putnam's theory. Sometimes trying to specify empirical findings numerically can throw into relief the fact that we have failed to ask some obvious questions about a case study. Equally, recourse to a case study can suggest the need for additional (or different) variables in a quantitative analysis. The two types of investigation should go hand-in-hand as each type can illuminate the other. Franklin (2004) uses both approaches in tandem in this way (see also note 7 below).

5 The ratio of each coefficient to its standard error is given in the column headed *t* in Table 13.8. This ratio determines the level of significance of each effect – the 'Prob' in Table 13.8 or the number of stars in Table 13.9.

6 Strictly speaking, an interaction term needs to be accompanied by both of the variables from which it is composed, and we would have retained the first election variable had the interaction term proved significant. But in small-*N* studies, this often is not feasible. We can justify eliminating one component of the interaction on the basis that the effect of the interaction goes up (from 9.41 in Model C to 10.92 in Model D) by the amount of the component that was eliminated (1.51). Technically, we prefer Model D for this reason rather than its higher variance explained. Model B (the alternative) does not account for both effects. (See Bramber, Clark and Golder, 2006.)

7 Those studying Switzerland never thought to consider that country's coalition arrangements as a source of turnout decline until a quantitative study (the Voter Turnout study mentioned in note 4) drew their attention to the likely importance of the so-called 'Golden Rule'.