PART II: LAWLIKE RELATIONSHIPS

The analysis of data is basically a matter of relating one variable to another. In effect we have already been doing this in Chapters 1 and 2 where we related the readings in one row to those in another, and in the last chapter where we related distance travelled to time taken, or consumers' expressed intentions-to-buy Brand X to whether they actually used it or not.

A relationship becomes *lawlike* when different sets of data are summarised or modelled by the same equation. Its status depends upon the range of empirical conditions under which it holds. This is discussed in Chapter 5. The uses of lawlike relationships are discussed in Chapter 6.

One particular use of a lawlike relationship is in analysing further data. However, sometimes we have to start without any previous result and in Chapter 7 we discuss the problem of fitting an equation to data for the first time. Often one fits a linear equation as a start. But many relationships are not that simple and Chapter 8 discusses the basic steps in dealing with non-linear relationships. Chapter 9 introduces problems where *many* variables have to be interrelated.

In the early stages, fitting a relationship is usually very empirical. Chapter 10 emphasises the emergence of *theory* with increasing knowledge of one's subject-matter.

CHAPTER 5

Descriptive Relationships

A relationship between two observed variables can often be represented by a simple mathematical equation. The primary criterion of such an equation is that it be *descriptive.* It has to summarise adequately the observed values taken by each of the variables.

The usefulness of a relationship lies in the range of conditions under which it holds. This notion of empirical generalisation is fundamental to science in general and to the study of relationships in particular.

5.1 Linear Relationships

The simplest relationship between two variables x and y is a straight-line equation of the form

$$y = ax + b.$$

Here a and b represent two numerical coefficients that stay constant for a particular relationship, such as

$$y = 0.5x + 2.$$

This equation says that for any given value of x, y is equal to 2 plus half the' value of x. Figure 5.1 illustrates this equation on a graph.

The quantity a is usually called the "slope-coefficient" because it measures the slope of the line when the equation is plotted on a graph. It shows that y varies by a units for every unit difference in x. The fundamental property of a linear equation is that the slope stays the same everywhere along the line.

The quantity b in the linear equation is often called the "intercept-coefficient" because it shows where the line intercepts the vertical (y) axis when $x = 0$. (This value may be well outside the range of the data and does not need to represent a physically relevant condition.) In our example, $b = 2$. A larger value of b would give a line parallel to that in Figure 5.1, but higher up.

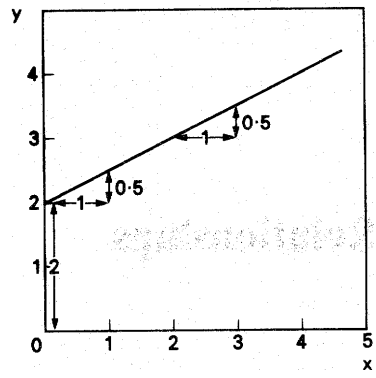

Figure 5.1 Equal Increases in y of 0.5 for a Unit Difference in x

When $b = 0$, the line goes through the origin, as shown in Figure 5.2A for $y = 0.5x$. Then y is *directly proportional* to x. When the coefficient $a = 0$, the equation $y = ax + b$ becomes $y = b$. This is illustrated in Figure 5.2B for $y = 2$. It means that the variables are independent of each other; i.e. y is *not* related to x and takes the same constant value of 2 irrespective of the value of x. This is the simplest possible result.

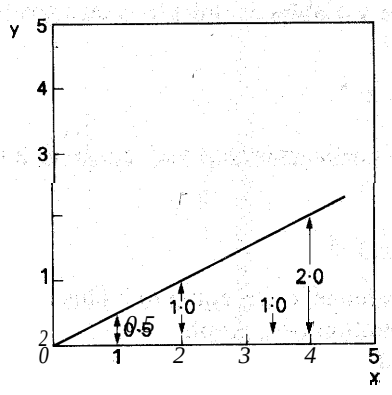

Figure 5.2A The Line $y = 0.5x$

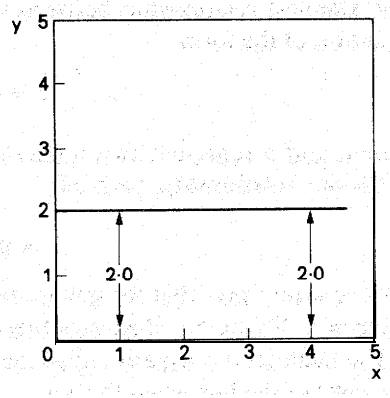

Figure 5.2B The Line $y = 2$

5.2 **Deviations ahd Scatter**

When a mathematical equation is used to describe a relationship between two variables x and y, the observed values do not generally lie exactly on the line. The line therefore is only an approximation to the observed data, which lie scattered around it. For this reason an equation like $y = ax + b$ should

strictly speaking be written as

$$Y \doteqdot ax + b,$$

where the symbol $\doteqdot$ means *approximately* equal and is perhaps the most important concept in applied mathematics. The equation could also be written

$$Y = ax + b \pm c.$$

Here the "plus or minus" symbol $\pm$ indicates that some observed readings have positive and others negative deviations from the theoretical line $y = ax + b$, and that these deviations have an average size of c units.

The existence of such deviations does not necessarily invalidate the relationship. Nor does the *size* of the scatter interfere with one's perceptual recognition of a relationship, as is illustrated by Figures 5.3A and B.

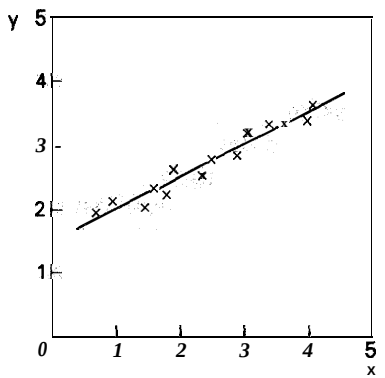

Figure 5.3A Small Scatter

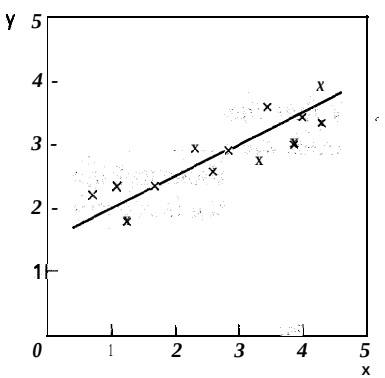

Figure 5.3B Large Scatter

The existence of scatter is generally taken for granted. Often it is not even explicitly described. Such an omission is open to purist criticism, but does not raise major practical problems. People are mainly concerned with knowing the *systematic* relationship between variables, e.g. that y varies as $0.5x + 2$, and not with the precise levels of scatter or error attached to it.

The crucial criterion is that the scatter should be *irregular*, i.e. show no systematic pattern, as in Figures 5.3A and B. Only then can it be "summarised away" in simple statistical terms, as being irregular (i.e. individually unpredictable) and of such and such an average size.

When the observed readings deviate *systematically* from the theoretical equation or line fitted, the deviations are more complicated to describe. This is the case irrespective of the size of the scatter. Figures 5.4A and B give examples of both relatively small and larger regular scatter. In Figure 5.4B we would have to say that when x is about $\frac{1}{2}$, y is almost $\frac{1}{2}$ a unit above the

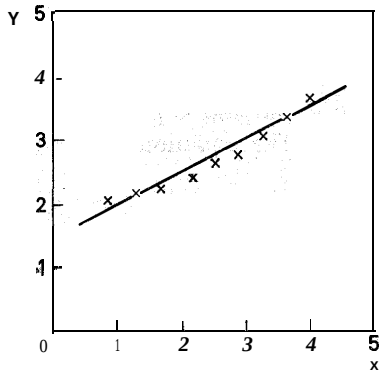

Figure 5.4A Small Systematic Deviations

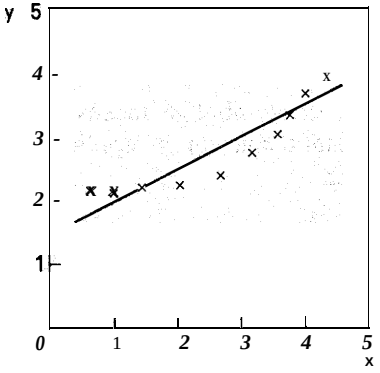

Figure 5.4B Large Systematic Deviations

line, when x = 1.5, y is on the line, when x increases to 2.5, y lies increasingly below the line, etc. The description of the data has become very complicated.

In the initial stages of data analysis, a systematic pattern in the deviations means that the wrong descriptive relationship has been fitted, even if the deviations are small. For example, although the deviations in Figure 5.3B are larger than those in Figure 5.4A, in the first case the scatter is irregular and in the second it is systematic. Usually it is more convenient to model such data with a suitable curve that has irregular deviations rather than with a straight line that has systematic deviations. Only in more advanced work are systematic deviations sometimes acceptable as deliberate oversimplifications, as long as their nature is understood.

5.3 **Non-linear Relationships**

Many scientific relationships and laws are highly non-linear in their original units of measurement. Boyle's Law in physics is a typical example. It says that the relationship between the pressure P and the volume V of a body of gas is

$$P = \frac{C}{V},$$

where C is a constant that depends on the amount and type of gas, the temperature, etc. This law states that for any given body of gas, the greater the volume, the less the pressure, but in such a way that the product PV remains approximately constant at C, or

$$PV = c.$$

The equation is non-linear because every unit change in V produces a *different* change in P. Figure 5.5A illustrates this relationship when C takes the value 10. The relationship between P and V is *different* everywhere along the line.

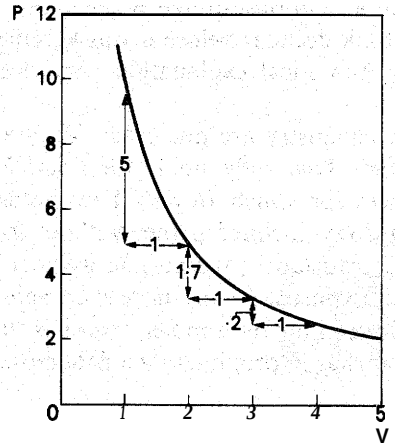

Figure 5.5A Boyle's Law: $P = C/V$

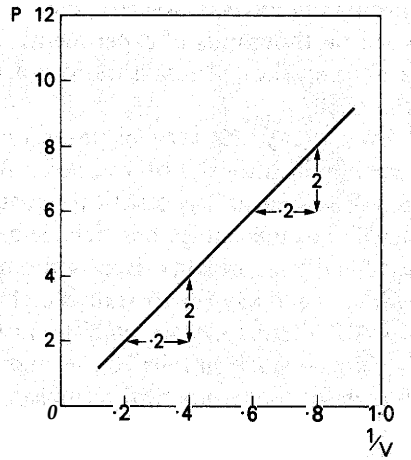

Figure 5.5B Plotting P against $1/V$

Many curved relationships can be described with a linear equation by changing the form of one of the variables. For example, with Boyle's Law we can "transform" the volume variable to l/V (the reciprocal of V) and then analyse the relationship

$$P = C\left(\frac{1}{V}\right).$$

This is the same as writing the linear equation $y = ax$, where y stands for P and x stands for l/V. Figure 5.5B shows that this transformed relationship produces a straight line where P varies by 2 units for every 0.2 difference in l/V-i.e. $P = 10(1/V)$.

Transformations like this are not always possible,. but they are worth doing because it is easier to fit a linear equation to transformed variables than to fit a curve to data in its original form,

5.4 The Status of Lawlike Relationships

A common comparison is between laws in physics and social science. Physical laws often appear to be exact, well-embedded in theory, and capable of giving simple, accurate predictions. By contrast, relationships in the

social and biological sciences seem merely to be statistical associations without any background of established theory, and capable of leading only to very uncertain and erratic predictions.

Such comparisons usually pit a well-established law from physics against some recent result from sociology, economics, or biology that may be based on only one or two isolated studies. But a relationship like Boyle's Law is based on thousands of experiments. It took decades before it was accepted as an empirical generalisation and its theoretical explanation came long after that.

Even today, the laws of physics and chemistry are not exact. They are oversimplifications. For example, Boyle's Law only holds for "perfect" gases (defined as hypothetical substances for which Boyle's Law holds); Newtonian mechanics has balls sliding down inclined planes without friction; Galileo's weights drop without air resistance; the atomic weights of hydrogen and oxygen are said to be 1 and 16 (instead of the more exact values of 1.00797 and 15.9994); etc. The justification for the simpler results is that they are so much simpler to use, and that they approximate the data closely enough for many practical purposes.

Thus lawlike relationships are not necessarily 100% exact. Nor do they *initially* have to form part of an explanatory theory or have an immediate practical use. It is usual for explanation, understanding and practical application to follow subsequently, as part of the historical development of a subject.

But this process cannot start until a relationship has first been discovered and described. We cannot explain *why* pressure varies with volume unless we know *how* it does so, e.g. as $P \doteqdot C/V$.

Such a descriptive relationship has to be a *generalisable* one. If for the next set of data the quite different relationship $P = -V$ were found instead, we would have to explain not only why pressure varies inversely with volume but also why it does so quite differently from one case to another. In contrast, explanation becomes relatively easy if the *same* relationship is found to hold in a variety of different circumstances.

5.5 Empirical Generalisation

The crucial step is establishing that the same quantitative relationship holds for different sets of data and different conditions of observation. Then the relationship becomes practical and useful.

In effect, this is what the laws of physics are: empirical regularities that have been laboriously isolated for a certain range of conditions of observation and that are equally well-known *not* to hold under other conditions. For example, Boyle's Law $PV \doteqdot C$ has been found to hold for different gases and mixtures of gas, different amounts of gas, different kinds of apparatus, different experimenters, different times and places, when pressure is *increasing*

and when pressure is *decreasing*, etc. But this was established only because of a massive amount of empirical observation that something like it in fact happened here and there, this year and last year, in the morning and at night, etc.

Empirical observation also showed that the relationship $PV = C$ does *not* hold when the temperature changes, when there is a chemical reaction, when there is a leak in the apparatus, when there is physical absorption or condensation of the gas, or when we tried to prove the law at school. Each time the law did not hold extensive empirical study was needed to generalise the relevant conditions. Was a particular deviation due to the specific gas being examined? A trace of water vapour that condensed? A careless laboratory assistant? Could the deviation be repeated, or was the situation not even well enough understood to do that?

The direct meaning and usefulness of Boyle's Law is therefore bound up both in our knowledge of the conditions under which it holds and the conditions under which it does *not* hold. The same process of empirical generalisation applies to any lawlike relationship. A recent, less-developed case is the study of how children's heights vary with their weights. The form of such a relationship might itself vary with a whole host of other factors like race, nationality, socio-economic class, sex, nutrition, age, point-in-time, etc. Yet it has been found that the same equation

$$\log w = .02h + .76$$

between the logarithm of children's weights w (in lbs) and their heights h (in inches) holds *despite* these differences, as listed in Table 5.1.

TABLE 5.1 Summary of Conditions Under Which the Height/Weight Relationship $\log w = .02h + .76$ Has So Far Been Found to Hold

(Lovell 1972, Kpedekpo 1970, 1971, Ehrenberg 1968)

Race:	White, Black, Chinese (in the W. Indies).
Countries:	U.K., Ghana, Katanga, West Indies, France, Canada.
Time:	1880 - 1970 approximately.
Age:	2 - 18 years.
Sex:	Male (2 - 18) and Female (2 - 13).
Socio-Economic Class:	Various in U. K., France and Canada.

The relationship also holds despite other less explicit differences in the conditions of observation, such as measurements being made by different observers and differences in the average size of families, housing conditions, intelligence levels, etc. No matter what differences there were, we know they did not matter because the same relationship has been found to hold.

The conditions covered in Table 5.1 look somewhat haphazard, like "Chinese children living in the West Indies". This is because the investigation

is at an early stage and no comprehensive or systematic checks of conditions have yet been made. Nevertheless, the range of conditions is already so wide that one may already begin to refer to the equation log w= .02h + 0.76 as a "lawlike relationship". One would now be surprised if the relationship did not hold for the next set of data from white or black children.

This does not mean that the relationship holds universally. There is always a variety of conditions under which a scientific law does ***not*** hold. Often this is so obvious that we automatically exclude such conditions even when thinking about the law. For example, the relationship between children's heights and weights obviously will not hold if they are measured when sitting or lying down, or for children who are seriously undernourished -we already *know* they will be light for their height. Less extreme exceptions are that teenage girls are heavy for their height compared with boys (which fits in with general experience of girls), and that babies may be relatively light for their length (babies being measured in the prone position).

5.6 **Other Things NOT Equal**

An empirical generalisation tells us how the values of the variables relate together despite variation in other factors. For example, the relationship log w = 0.02h + 0.76 between heights and weights holds despite differences in children's races, nationalities, sex, ages, social classes, point-in-time, etc. Similarly, in Part I we saw that the quarterly readings in the North were about 95 even though Quarter I was Winter, with ice and snow and long dark nights, Quarter II was Spring with April showers, apple blossoms and lambs gambolling, and so on.

The popular saying "other things being equal" therefore does not necessarily mean that all these other things have themselves to be equal. It only means that their *effects* have to be equal.

5.7 Summary

A relationship between two variables usually can be represented by a mathematical equation. The prime purpose of this equation is to be descriptive: to summarise in a convenient form the observed values taken by one variable for any given value of the other variable.

The simplest form of relationship is the straight line equation $y = ax + b$, which says that y varies with x at a constant rate. In practice most empirical relationships are non-linear in their original units of measurement. However, one variable can often be "transformed" to reduce a curved relationship to a linear form.

Observed readings are usually scattered about a theoretical equation or fitted line instead of lying exactly on it. Deviations that have no regular

pattern of their own are easy to summarise statistically as being irregular and of a stated average size.

To be lawlike, a relationship has to be based on many different sets of data and hence generalise to a wide range of different conditions of observation. This does not imply that the relationship holds universally, but only that it holds within the stated range of conditions and that the exceptions can also be generalised.

CHAPTER 5 EXERCISES

Exercise 5A. What is a Variable?

A common dialogue about variables runs as follows:

Teacher: "Suppose that x is a variable, i.e. a quantity which can take any value."
Student: "Yes, I think I understand that."
Teacher: "Let x = 20."
Student: "But you just said x was a variable!"

At this stage, many students are lost for good. But is there really a contradiction in what has been said?

Discussion.

In algebra, symbols like x are used to represent any value that a particular quantity can take-usually within certain stated or implied limits, e.g. that x is always a positive integer (as in counting $0, 1, 2, 3, 4, \ldots$), or that x varies only from 0 to 1 (e.g. a proportion, .3, .6, .1, etc.).

If x is a variable, it can therefore take various possible values. Saying that x = 20 is a way of looking at one of these values. It presents no contradiction.

Although we shall be using almost no complex notation in this book, it is worth briefly illustrating some common elaborations. For example, more detailed notation is sometimes used to distinguish x as a general variable from x as a particular value. We might use the symbol x' (usually called "x dash") for a particular value and say that x' = 20. Even here, x' still represents a variable quantity: it could be *any* particular value of x. But for the moment we are saying that we are considering the case where x' takes the value of 20.

Introducing such additional notation is especially useful when we wish to say more about some particular value of x. Suppose we want to consider two values of x which are not equal. We can write this as x' ≠ x", where the symbol ≠ means "not equal", and x" is another particular value of x (called "x double dash"). We could not write this without some notation distinguishing the two values of x.

Or we might want to consider all those pairs of values of x which differ by 5 units. These we could denote by the equation x" = x' + 5. Again we are referring to all possible values of the variable x which satisfy the condition laid down by the equation.

The values x' and x" in the two equations x" ≠ x' and x" = x' + 5 do not necessarily refer to the same possible values of x. For example, the values

x' = 20 and x" = 21 would satisfy the first equation but not the second. If we wanted to differentiate the two cases symbolically, we would have to introduce a further difference in notation, like x' ≠ x" and $x_2 = x_1 + 5$.

Using such different symbols can become cumbersome. When no real confusion can arise it is therefore common to use the simple symbol x to stand for both the variable itself and some specific value of it. While mathematics is generally a very precise subject, the symbolism is often used very loosely to keep some flexibility and simplicity.

Exercise 5B. Algebraic Relationships

In the linear equation y = ax + b, x and y are two variables and a and b are two constants that can take any values. Why are a and b called "constants" and not considered variables as well?

Discussion,

When we are speaking about linear equations *in general we* denote them by y = ax + b. But whenever we speak of a particular linear equation, the coefficients a and b automatically take numerical values that remain constant for that specific equation. On the other hand, x and y are always variables in all linear equations.

For example, we can speak of the linear equations

$$y = 0.5x + 2,$$

$$y = 2x + 6.$$

In each equation a and b take specific constant values and x and y are variables.

Exercise 5C. Points on a Straight Line

On a graph of y and x, the values (x, y) refer to the point which is x units measured along the x-axis and y units along the y-axis. If (x_1, y_1) and (x_2, y_2) are two points that lie on the straight line $y = ax + b$, what can we say about the relationships between the four values x_1, x_2, y_1 and y_2?

Discussion.

Since the point (x_1, y_1) lies on the straight line, it must satisfy the equation

$$y_1 = ax_1 + b.$$

Similarly, for (x_2, y_2) we must have

$$y_2 = ax_2 + b.$$

Subtracting one equation from the other, we have

$$\begin{aligned} y_1 - y_2 &= ax_1 - ax_2 \\ &= a(x_1 - x_2). \end{aligned}$$

This says that for any two points that lie on the straight line, the difference between the two y-values $(y_1 - y_2)$ is always a times the difference between the two corresponding x-values $(x_1 - x_2)$.

It follows that two points determine a particular straight line. Thus from the above equation we have that

$$\frac{y_1 - y_2}{x_1 - x_2} = a$$

This allows us to calculate the slope, a, of that line. For example, for the two points $(2, 3)$ and $(4, 4)$ we have

$$a = \frac{3 - 4}{2 - 4} = \frac{-1}{-2} = 0.5.$$

The slope-coefficient $a = 0.5$ will be the same for any two points on that line, which is the fundamental property of the straight line.

The intercept-coefficient b can be calculated by noting that the point $(2, 3)$ is supposed to lie on the line. Thus

$$3 = 0.5(2) + b$$

and so $b = 2$. The linear equation is therefore

$$y = 0.5x + 2.$$

(The same result could be obtained using any point, e.g. $(4, 4)$, on that line.)

This is the normal way of determining the numerical values of the coefficients a and b, given two points which lie on the straight line.

Exercise 5D. A Change in Units

If $d = 0.5t + 2$ is an equation between distance d in feet and time t in seconds, what is the corresponding equation using miles and hours?

Discussion.

There are 5,280 feet in one mile and 3,600 seconds in one hour. Thus the distance D in miles is $D = d/5{,}280$ and the time T in hours is $T = t/3{,}600$. (D and T must be numerically smaller than d and t because they are measured in larger units.) Substituting for d and t in the equation $d = 0.5t + 2$, we therefore have

$$5{,}280D = 0.5(3{,}600T) + 2.$$

This gives

$$D = .34T + .00038.$$

Exercise 5E. x in terms of y

If $y = 0.5x + 2$, what is x in terms of y?

Discussion.

Isolating the x-term gives $0.5x = y - 2$. Dividing the equation by 0.5 (or multiplying by 2), we have

$$x = 2y - 4.$$

In general, if $y = ax + b$, then $x = (y - b)/a$.

Exercise 5F. Scatter about the Equation

If the observed reading (x', y') does not lie exactly on the theoretical line $y = ax + b$, what is the deviation?

Discussion.

Given the line $y=ax+b$ and the value x', the theoretical y-value on the line is $ax'+b$. Hence the difference between the theoretical and observed values of y' is

$$y' - (ax' + b), \text{ or } y' - ax' - b.$$

If the point is (4, 3) and the line is $y=0.5x+2$, the deviation measured on the y-scale is

$$3 - 0.5(4) - 2 = -1,$$

as shown in Figure 5.6.

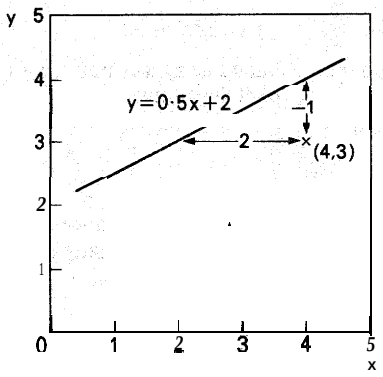

Figure 5.6 Deviations in the y and x Directions from $y=0.5x+2$

Measured on the x-scale, the deviation between the observed and theoretical values of x' is

$$x' - (y' - b)/a = x' - (y' - 2)/0.5.$$

For the point (4, 3) this is $4 - (3-2)/0.5 = 2$, as is also shown in Figure 5.6.

In other words, the point (4, 3) is 1 y-unit too low or 2 x-units too high, compared with the line.

CHAPTER 6

Using a Given Relationship

Lawlike relationships can have many practical uses. The illustrations in this chapter are based largely on the relationship between the height and weight of children referred to in Section 5.5. This case is typical of what one meets in practice, since it is non-linear, new, and incomplete.

6.1 Summarising the Available Data

The most direct function of an empirically-based relationship is to have summarised the data on which it is based. For example, the height/weight relationship, $\log w \doteqdot .02h + .76$, summarises the heights and weights of certain British boys aged 5 in 1880, of West Indian Chinese boys of various ages in the 1960's, of French girls aged $8\frac{1}{2}$ in the 1950's, and so on. Once one has this relationship one need not refer again to the raw data. That is a major achievement.

6.2 Prediction and Extrapolation

A relationship like $\log w = .02h + .76$ can also be used to predict the value of one variable from the known value of the other. In effect one is then asserting that the relationship with its associated scatter will hold again for a new set of children. That is what one expects to happen if the new data fall within the range of conditions already covered in the previous analyses.

However, one may sometimes need to extrapolate to new conditions, outside the range previously covered; e.g. for some very different racial or ethnic group, for children living under unusual circumstances or suffering from some illness, for those who lived 300 or more years ago, for another species, or whatever.

The difference between these two uses turns on the extent to which the new data lie within the range of conditions already covered. 'The distinction between a well-founded prediction and an extrapolatory guess can be seen by considering the actual outcome in each case: i.e. either success or failure.

With a prediction, success would be just what one had to expect. All the previous experience has been that log w $= .02h + .76$ holds under such conditions and now it has done so again; a dull but comfortable outcome. In contrast, *failure* of such a well-founded prediction signifies a discrepancy with all the previous knowledge, something new and potentially interesting which must be studied further and explained.

But with an extrapolatory guess, it is *success* that represents an exciting discovery, a further generalisation of the relationship log w $= .02h + .76$ to quite new conditions. *Failure* merely means that the analyst is not very good at making such guesses; there is no actual discrepancy because nothing was previously known about the relationship under these conditions (e.g. for the heights and weights of young chimpanzees).

6.3 **Understanding and Theory**

A well-established relationship also allows us to reach a better understanding of the phenomena in question. It might be thought that a purely descriptive generalisation like log w $= .02h + .76$ only shows *how* height and weight are related, but does not tell us *why*. But this is not entirely true. Consider all the factors which might affect the way height varies with weight: race, nationality, sex, social class, age, time, observers, etc. We now know that these factors generally do not affect the relationship. We also now know that some variables *do* affect the height/weight pattern, such as puberty in girls. Clearly we are beginning to understand something about the system.

Nonetheless, the relationship is still a limited result. For example, it hardly links up with other kinds of findings. The height/weight relationship is in fact lacking in "theory". This is essentially a matter of time because the relationship is still relatively new. A generalisable result has first to be *established* before it can be incorporated into any broader system of equations or theory. Low-level empirical generalisations are the essential building-bricks of more advanced theory; examples of such extensions are discussed later, particularly in Chapters 9 and 10.

6.4 **Technological Applications**

A technological use of the height/weight relationship is to assess whether an individual child is above average weight, as a step toward diagnosing possible obesity.

Suppose a particular child is 51 inches tall and weighs 63 lbs: which from a table of logarithms is about 1.80 in log lb units. These values do not satisfy the equation log $w = .02h + .76$ exactly. For a value of $h = 51$, the equation gives $.02h \times 51 + .76 =$ **1.78.** This differs by .02 log lbs from the observed weight.

The child is therefore "overweight" in the sense that the equation gives the *average* result and the child is heavier for its height than average. None of the sources quoted in Table 5.3 reported individual deviations from the equation, but from some data supplied by Dr. E. M. B. Clements (1954), it appears that the average size of the deviations of individual children might be about .04 log lbs. A child who is .02 log lbs above average therefore is not *abnormally* overweight. There are many children of that height who are even heavier (i.e. who differ even more from the equation).

It may seem difficult to think of many other technological applications for the height/weight relationship. But this is usual for a *single* relationship in its early days. For example, by itself the law of gravity does not tell us how to build an aeroplane that will fly or to adjust a pendulum clock to give the correct time, The law is just one of many that engineers use in these cases.

6.5 **Decision-making**

In using data for *decision-making*, such as whether to introduce free milk at school for children from less prosperous backgrounds, or deciding whether to believe that black children are of a different shape from white ones, it is best to have first summarised and understood the available data. Decision-making and data-analysis are separate processes.

6.6 The Analysis of Further Data

Another use of a relationship is in dealing with further data. This makes the new analysis very simple. One merely checks whether the previous result holds again. For example, in Section 6.4 we very easily concluded that the child's weight was well within the usual limits.

Use of previous results helps keep the analysis simple even when the new data are extensive. Table 6.1 gives an extract of the average heights and weights of about 5,000 children in Ghana (Kpedekpo, 1970). They are classified by sex, yearly age-groups, and race, with the black children further sub-classified as living under rural, urban non-privileged and urban privileged conditions (i.e. attending élite schools).

The biggest variation in Table 6.1 is between age-groups. We therefore start by analysing a particular type of child at different ages and in Table 6.2 compare the expatriate boys with the earlier relationship. The fit is clearly good. The corresponding analyses for the expatriate girls and the privileged black children show that they also follow this relationship. The deviations are within the same average limits of $\pm .01$ log lb that were reported for such age-groups in earlier studies (Ehrenberg, 1968). The slightly larger

TABLE 6.1 Average Heights and Weights of Groups of Children in Ghana

(From Kpedekpo 1970)

HEIGHT (in inches)	Age in Years 5	6	7	8	9	10	etc.
BOYS							
Rural	*	45	47	49	51	.	.
Urban	42	45	47	48	50	.	.
Urban privileged	45	48	49	51	54	.	.
Expatriate (white)	45	46	49	52	54	.	.
GIRLS							
Rural	44	45	47	50	51	.	.
Urban	*	46	48	50	51	.	.
Urban privileged	45	47	48	51	54	.	.
Expatriate (white)	45	46	49	52	53	.	.

WEIGHT (in lbs.)	Age in Years 5	6	7	8	9	10	etc.
BOYS							
Rural	*	41	45	51	55	.	.
Urban	35	41	44	48	54	.	.
Urban privileged	46	52	58	62	69	.	.
Expatriate (white)	46	50	59	66	69	.	.
GIRLS							
Rural	41	41	45	52	54	.	.
Urban	*	46	50	52	57	.	.
Urban privileged	46	50	59	64	73	.	.
Expatriate (white)	44	49	57	66	67	.	.

* No data

TABLE 6.2 Expatriate Anglo-American Boys in Ghana and the Prior Relationship log w = .020h + .76

Expatriate Boys in Ghana	Age 5	6	7	8	9	10	etc.	Av. (5 to 9)
Av. height : h	45	46	49	52	54	.	.	49
Av. weight : log w	1.65	1.70	1.76	1.82	1.84	.	.	1.75
.020h + .76	1.66	1.68	1.74	1.80	1.84	.	.	1.74
log w - (.020h + .76)	-.01	.02	.02	.02	.00	-	.	.01*

*Average size ignoring sign = .01

deviations in the table at .02 log lbs do not generalise (e.g. to older Ghanaian boys, girls, etc.).

But analyses of the black non-privileged urban and rural children show consistent negative deviations from the relationship, as typified in Table 6.3. They are consistently lighter by about .04 log lbs for any given height than white or "privileged" black children. (The apparent trend in the deviations in Table 6.3 does not generalise.) Here use of the prior relationship makes it clear what new relationship will describe these discrepant data: something like $\log w = .02h + .72$.

TABLE 6.3 Non-Privileged Urban Ghanaian Boys and the Prior Relationship log w = .020h + .76

Non-privileged Ghanaian Boys	Age 5	6	7	8	9	10	etc.	Av. (5 to 9)
Av. height: h	42	45	47	48	50	.	.	46.4
Av. weight: log w	1.54	1.61	1.64	1.68	1.73	.	.	1.64
.020h + .76	1.60	1.66	1.70	1.72	1.76	.	.	1.69
log w - .020h - .76	-.06	-.05	-.06	-.04	-.03	.	.	-.05

These analyses have clearly been very easy to do, regardless of whether they showed agreement or disagreement, because we could use the prior relationship.

6.7 The Meaning of Failure

Using a prior relationship to analyse new data therefore works even if the new data do not agree with the earlier relationship. This is illustrated in Figure 6.1, where the previously established line clearly does not hold for the new data. Such a result may spell failure to what one hoped to find, but technically it is very simple: there is no generalisable relationship. One variable cannot be predicted from the other, at least not until the discrepancy is itself explained. But this may not be possible and such unresolved failure is worth illustrating.

Some years ago, a series of experiments was mounted to see whether a certain chemical measure could be used to predict the "eating quality" of white fish (cod, haddock, etc.). The practical relevance was that while the sensory assessment method was known to be highly reliable, it was not very acceptable to trawler captains for price-setting purposes, nor was it easy to operate routinely at 6 am each morning at the quay-side. An "objective"

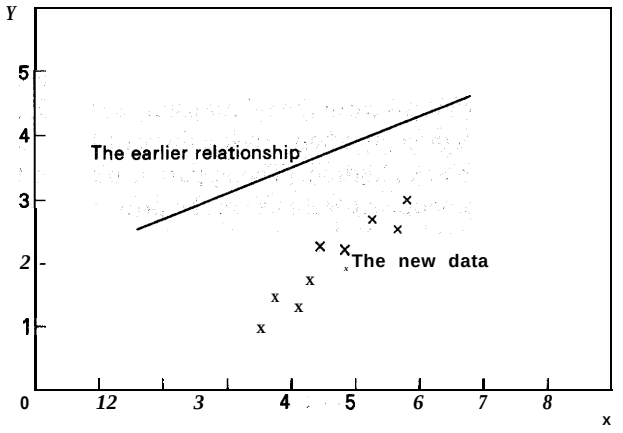

Figure 6.1 New Data which is Different

chemical measure of eating quality was therefore desirable. The two variables were

(i) V, the amount of volatile bases per milligram of fish muscle (i.e. chemical compounds such as ammonia and various amines);

(ii) F, the flavour of the fish, a direct indicator of its "eating quality" as measured on a lo-point scale by a highly trained laboratory taste panel.

In some pilot studies in 1953, the relationship $F = -6.2 \log (1+V) + 15.0$ had been found to hold between F and V for batches of fish stored in ice for various periods (Shewan and Ehrenberg, 1955). It therefore seemed that one variable could be used to predict the other. However, the relationship had not yet been established for the range of conditions that existed in practice: fish from a variety of fishing-grounds, caught at different seasons of the year, handled and stored in different ways on board trawlers, etc. As a first follow-up, further measurements were made for several hundred batches of fish caught at different seasons of the year in 1954 and 1955 (Shewan and Ehrenberg, 1957).

Each new experiment gave well-fitting relationships of the form $F = a \log (1+V) + b$, but the coefficients a and b differed markedly each time, and therefore also from the initial equation. For example, in one case the relationship was something like $F = -12 \log (1+V) + 20$, in another $F = -4 \log (1+V) + 10$, in a third different again, and so on, as illustrated graphically in Figure 6.2.

No explanations for these discrepancies were found. Possible factors studied included the different seasons of the year, the size and sexual maturity of the fish, differences in the nature and size of the initial bacterial load of the fish, variations in the chemical composition of the fish muscle, differences

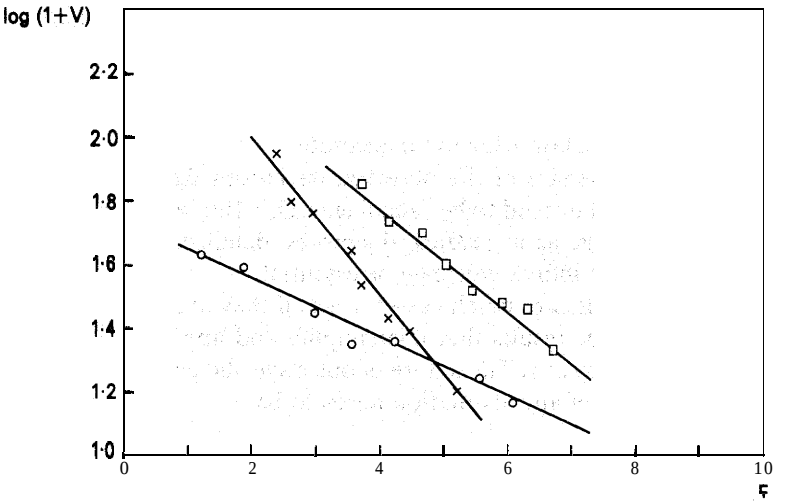

Figure 6.2 Irreconcilable Equations between the Volatile Base Content V and the Flavour F of Cod from Three Experiments

in handling or storage methods, and the influence of errors of measurement (which were small and well-understood for both measuring techniques). But whatever additional factors were involved in the relationship between F and V, they remained completely unknown. Neither the initial relationship nor any of the subsequent results were reproducible. The chemical measure might appear more "objective", but the analysis showed that it objectively failed to measure eating quality in a reproducible way.

6.8 The Purpose of the Analysis

The purpose of an analysis does not affect how it is done. For example, a result used for prediction is of the same form as one used to analyse new data. In all cases one has to describe how y varies with x within the observed range of conditions. One's purpose, however, influences *other* things, such as whether to do the analysis at all, how much effort to put into it, which kinds of deviations to follow up, what range of conditions to cover, how much accuracy or precision to try for, and so on.

In the white fish study for instance, it might be enough for some purposes to know that flavour and volatile bases content are always inversely related, the more volatile bases in the fish, the lower its flavour-score, even if the numerical details vary from case to case. But for quality-control and price-setting purposes the unpredictable variations in the numerical coefficients are too large. Far too many fish of acceptable flavour would be wrongly

given a low price (suitable for fish-meal, say) and too much spoilt fish awarded a high price.

Again, for many purposes the pressure and volume of a gas can be taken to follow Boyle's Law, $PV \doteqdot C$. But where gases are under high pressures, as in oil refineries, this law is far too inaccurate.

More generally, the size of the observed deviations from a law is often irrelevant as long as they tend to be "small enough". But in making decisions about individual cases, as in medical diagnoses, detailed understanding of the variation between individuals may be essential.

Since analytic results are much easier to use if they are relatively simple, we generally strive for results that oversimplify and approximate the data rather than fit them exactly. The nature of one's specific problem determines how close the degree of approximation needs to be.

6.9 **A Common Misuse**

Relationships are often misused by applying them unthinkingly outside the range of conditions for which they have been established. If the equation $y = 5x + 10$ fits certain data, it is often assumed that if x is changed by a certain amount, then y should correspondingly change by 5 times that amount. But this is not necessarily true.

The initial data may not have referred to *changes* in x and *y*, and the relationship therefore cannot tell us anything directly about such changes. For example, the earlier height/weight relationship describes' how the heights and weights of many different kinds of children are related. It does not say that any child will grow taller if one increases his weight by feeding him a lot, or that he gets shorter whenever he loses weight. Even for two variables like weight and *girth*, where we know from everyday experience that short-

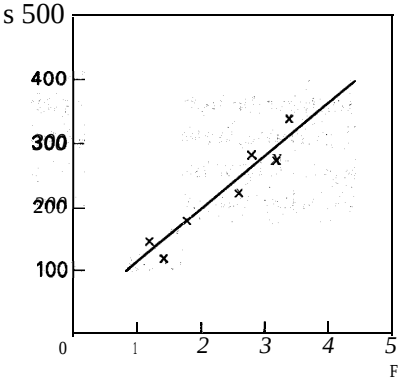

Figure 6.3 Floor-Space *F* and Sales-Level *S* of Some Different Retail Stores

term ups-and-downs do tend to correlate, it would be wrong to assume that these short-term changes necessarily follow the same quantitative pattern as those which occur as children grow.

As another example, suppose there is a close relationship, $S \doteq 80F + 41$, between the sales-level S (in thousand of pounds sterling) and the floor-space F (in thousands of square feet) of a number of retail shops, as illustrated in Figure 6.3. It does not follow from this that increasing the floor-space of a shop will lead to a corresponding increase in sales.

Variations among different shops and changes within a particular shop are two different things. For example, we know that there can be seasonal increases in sales with no change in floor-space, as shown in Figure 6.4A. This is quite unlike the relationship $S = 80F + 41$. Or all the shops could be rearranged to increase their effective floor-space by about 25 %, but this might have no effect on sales, as Figure 6.4B illustrates.

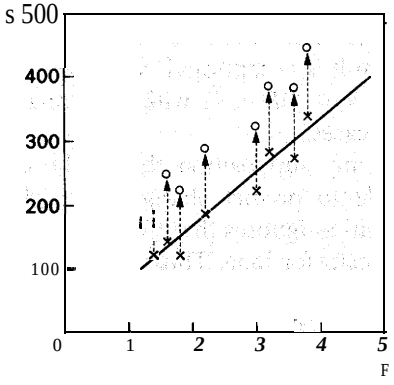

Figure 6.4A A General Increase in Sales

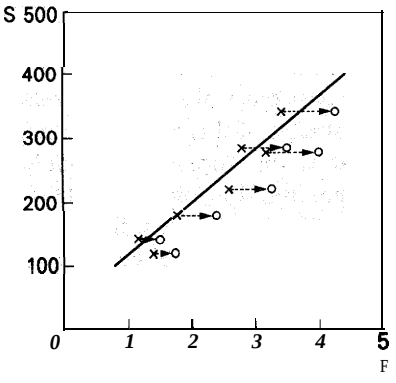

Figure 6.4B A General Increase in Floor-Space

Many relationships do not tell us directly what we want to know for practical decision-making purposes. The decisions may be concerned with some deliberate man-induced change occurring over time, the given relationship with existing, static and cross-sectional differences. (Much of economic analysis, for example, ignores this distinction.) Nevertheless, a static cross-sectional relationship tells us about certain constraints in the system which we may be trying to change. The equation $S = 80F + 41$ shows how S and F are related for the different shops within the conditions covered by the data. It does not say what will happen to one variable if one deliberately changes the other variable, but it does tell us about the context in which any change would take place.

Suppose for instance that the floor-space of a particular shop is doubled (e.g. rebuilding, or taking over the next-door premises), but sales have

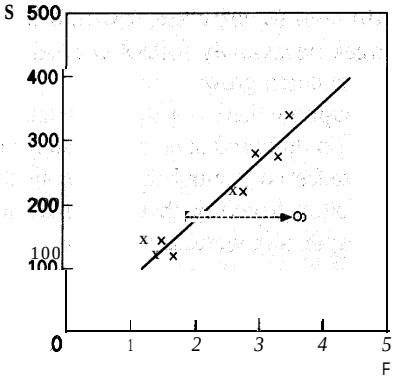

Figure 6.5 Doubling the Floor-Space of a Particular Shop

remained virtually the same, as shown in Figure 6.5. Clearly we have a clash with the equation $S = 80F + 41$. The new result lies well outside the previous range of scatter for different shops. The result is exceptional; we would not in the past have observed the relationship $S = 80F + 41$ with its relatively small scatter, if there had been many such cases:

That is the apparent rationale behind any supposition that the cross-sectional relationship $S = 80F + 41$ *ought* to predict change, i.e. that if F increases, S must also. But such a supposition ignores that the exceptional situation depicted in Figure 6.5 need not occur for long. Thus

(i) Something may have been done to increase sales of the shop *after* the increase in floor-space (e.g. publicity), thus bringing it into line again with the norm $S \doteqdot 80F + 41$.

(ii) Floor-space may have been reduced or used for other purposes after a while because sales did not warrant the extra space.

(iii) The shop may have gone bankrupt or been closed because it was losing money, so that it would disappear completely from the system modelled by the equation.

(iv) The shop may have continued with its abnormally large floor area with the extra costs covered by higher retail margins. Other such cases might have been deliberately excluded in the past from the initial analysis because they were atypical of the type of shop being studied!

(v) The shop may remain as the exception to prove the rule, the one case where a manager doubled his floor-space and did not increase his sales: an "awful warning" so well-known that it stopped repetition.

(vi) Maybe no such cases have ever been observed, anyway!

Any attempt to interpret the cross-sectional relationship $S = 80F + 41$, as telling us that sales will increase if we increase the floor-space ignores

all such possibilities. It also ignores the possibility that if a causal chain exists, it may be in the *reverse* direction, with increases in floor-area being decided on after sales *have* increased (or are "certain" to do so because a new housing development is nearing completion).

6.10 Simple Prediction Again

The complexities in the last section arose because of extreme extrapolations beyond the range of conditions covered by the observed relationship, $S = 80F + 41$. The equation only approximated how sales and floor-space varied together among the different shops in Figure 4.3. It might therefore be thought that the relationship could be used more easily in its own context, e.g. to predict that the sales of an additional shop with floor-space F' should be about $(80F' + 41)$. But there are some difficulties even in that. For example, the sales of the new shop will obviously be nothing like that unless someone remembers to put in stock, hire staff, and unlock the doors at the right times. More generally, the equation $S = 80F + 41$ may be relevant only if the shop is managed in the same style as the original shops.

The prediction for a particular shop will therefore be firmer if the original equation held for shops managed in various *different* ways. Then we would know that certain variations in management style do not affect the relationship between S and F, and it does not matter precisely how the new shop is to be run, as long as it is within the range already covered. Similarly, we need to know the type of locations of the seven shops, what kinds of shops they were (supermarkets, corner-shops etc.), the ages of the premises, and so on. In particular, if the previous data referred to *established* shops and the additional shop is altogether new and not based on a previous business in the same location, presumably it will take some time for its sales to build to a "normal" level.

Thus even the simple predictive use of an equation like $S = 80F + 41$ is not really that simple. It does not depend so much on statistical matters, such as the number of readings on which the relationship is based, or how closely the equation fits these readings. Instead it depends on two quite different factors: the *range* of conditions under which a generalisation has been established, and whether the additional shop lies *within* that range of conditions.

6.11 The Need for Research

The need throughout is for empirical generalisation. This shows that trying to establish lawlike relationships is not for the amateur. One is unlikely to have much success by merely collecting some data and a statistical technique and applying one to the other. The problems are not necessarily

difficult, but they are laborious. Developing a well-based empirical law requires a great many different sets of data and much analytic effort and perseverance, mostly at a rather tedious level of detail.

For example, current doubts about the validity and practical applicability of the height/weight relationship rest only marginally on the nature of the analyses carried out so far. The real worry is whether the same relationship will hold under other kinds of conditions that might be met in the future. This worry can be reduced only by extending the range of conditions that have already been examined in the past. So far the results are based on four studies only, the earliest published in 1968. The authors were primarily concerned with data to which they had easy access. Although each case covered an extensive range of different sets of data no *comprehensive* studies have yet been undertaken. Thus the empirical basis of the relationship is still relatively weak. Typically, more cross-checking with other data is needed.

It follows that the potential user of a relationship-the engineer, doctor, economist, administrator or scientist-can seldom have instant answers to his problem, let alone expect to obtain valid do-it-yourself results from scratch. Instead, investment in research is needed that will provide well-established results for subsequent use. This takes time. The skill in managing the necessary research and development work lies in starting the right research at the right time.

6.12 **Summary**

The most immediate practical use of a lawlike relationship is that it reduces bulky data to a succinct summary or model. We can then predict that the relationship will hold for other cases within the range of conditions already covered. We can also *extrapolate* outside these conditions by making an informed guess.

Other uses of lawlike relationships lie in technological applications, in leading to a better understanding of the phenomena in question, and in providing the basis for more ambitious theories and explanations.

Lawlike relationships are misused when they are applied unthinkingly to conditions not covered by the previous data. Typically, the relationship $\log w = .02h + .76$ tells us how the height and weight of different children are related. It does not say how a *change* in the weight of a particular child will be related to a *change* in his height. The two situations are different and the relationships between the variables cannot be assumed to be the same. This is a question for *empirical* investigation and analysis.

A previously established relationship also greatly facilitates the analysis of new data. One need merely check the new data against the already available equation. This is easy to do. If the result is successful, it leads to

a wider generalisation of the relationship. But if the previous relationship and the new data fail to agree, this means the relationship does not generalise, and then one has no basis for prediction or for technological applications. This result is also worth knowing.

CHAPTER 6 EXERCISES

Exercise 6A. The Meaning of a Relationship

If the sales-volume S and floor-area F of different shops are related by the equation $S = 80F + 41$, does this mean that shops will *shrink* if sales drop?

Discussion.

This kind of example is a popular way of warning against the misinterpretation of equations by applying them to conditions for which they have not been validated. Unfortunately this lesson is then often ignored in less obviously unreasonable situations. For example, it is still assumed that increasing floor-area will increase sales. The point is that of course it *might* do so, but the "cross-sectional" relationship $S = 80F + 41$ as such does not provide any direct evidence that it would.

Similarly, decreasing sales *do* often cause shops to shrink, but not because of the equation $S = 80F + 41$. The reductions in floor-space are not even in line with it numerically. A drop in sales below a certain level may cause a shop to be closed completely, and we may regard a closed shop as effectively having no floor-space. But the result is quite unlike $S = 80F + 41$.

This also raises the practical question of how one's variables are operationally defined. For example, how does "floor-space" differentiate between selling space, storage space, space for staff amenities, etc? Before attempting any deep explanation or extrapolation of a relationship, it is wise to clarify what the variables mean.

Exercise 6B. Less Simple Data

The analysis in Section 6.6 of the height and weight data for children in Ghana was easy to do because we chose to look at a particular type of child (e.g. expatriate white boys, or non-privileged Ghanaian girls) across different age-groups. In each case this gave consistent results, either consistent agreement as in Table 6.2, or consistent disagreement as in Table 6.3. This was not accidental. We knew age was the biggest single factor and structured the analysis accordingly. But sometimes this prior knowledge can be misleading, and one does not always have it.

As an alternative approach to the data in Table 6.1, analyse each age-group separately. An example for the 9-year-olds is given in Table 6.4, where the prior equation, $\log w = .02h + .76$ is also fitted.

Discussion.

Table 6.4 for the 9-year-olds shows mixed results. Expatriate white and privileged black children are in line with the previous result

TABLE *6.4* Heights and Weights of 9-Year-Old Children in Ghana and the Prior Relationship log w = .02h + .76

(From Table 4.1)

9-Year-Olds	Sex and Living Conditions								
	Boys				Girls				
	Rur	Urb	Priv	Exp	Rur	Urb	Priv	Exp	Av.
h	51	50	54	54	51	51	54	53	52
log w	1.74	1.73	1.84	1.84	1.73	1.76	1.86	1.83	1.79
.02h + .76	1.78	1.76	1.84	1.84	1.78	1.78	1.84	1.82	1.80
Difference	-.04	-.03	.00	.00	-.05	-.02	.02	.01	-.01

log w = .02*h* + .76 ± .Ol. But rural and urban black boys and rural girls are substantially lighter for their height, by about .04 log lbs.

The result for the urban (non-privileged) girls is somewhat unclear. The deviation of −.02 is no more than the occasional large deviation found for the basic relationship. These girls could therefore be in line with this relationship. But they might also fit in with the other rural and urban 9-year-olds.

However, the urban (non-privileged) girls of *other* ages are generally .04 log lbs too light, as are all the urban boys and rural children. Thus the urban g-year-old girls fit in with *them.*

Another query is for the *privileged* black g-year-old girls. Other ages in this category are on average within .Ol log lb of the basic equation, Thus the somewhat large deviation of +.02 for this group in Table 6.4 seems to be merely the occasional larger deviation from zero.

This analysis is only a slightly clumsier way of reaching the same conclusion as in Section 6.6. But it is typical of the slower "teasing-out" of results that tends to be required in practice. One's first attempt to select sub-groupings of new data is not always the simplest to use.

Exercise 6C. The Fit for Sub-groups and Aggregates

If a relationship holds for a certain set of data, will it also hold for any sub-group?

Discussion.

If there are a number of groups of readings and a relationship holds for each of them, it must also hold for the aggregate or combined set of data. (Table 6.2 in Section 6.6 gives an example.)

But the opposite is not necessarily true. If there is scatter about an equation, any particular sub-group may consist of individual readings which are biased in one direction. This could "average out" when considering the *total* set of readings. Table 6.4 in the previous Exercise gives an example. The fit for *all* the 9-year-olds appears good, but the fit for the rural children is certainly not.

Exercise 6D. $y = ax + b$ or $x = y/a - b/a$?

In Chapter 4 we used the equation $I = K\sqrt{U}$ between intention-to-buy I and usage U. Could this also be written as

$$I^2 = K^2U, \text{ or}$$

$$U = I^2/K^2, \text{ or}$$

$$u = LI^2, \text{ where } L = 1/K^2?$$

Discussion.

The three equations are mathematically identical. The formulation $U = I^2/K^2$ would be used to determine the value of U corresponding to a given value of I, for example when predicting U from I.

But for the general reporting of results, the form $I = K\sqrt{U}$ is preferable because the scatter of the data is easier to summarise in this form. The deviations $(I - K\sqrt{U})$ have a constant average size of about $\pm$ 3 percentage points all along the line: i.e. they are "homoscedastic". It follows that the average size of the deviations $(I^2 - K^2U)$ about the line $I^2 = K^2U$ will *not* be constant: they will differ for low and high values of U. Similarly, the scatter for the formulation $U = I^2/K^2$ will not be constant. We illustrate this for the equation $I = 10\sqrt{U} \pm 3$, where $K = 10$.

Consider first a relatively low value of U, say $U = 4$. For this, I takes the value $10\sqrt{4} = 20$. The average limits of scatter of I will be from 17 to 23. In the formulation $U = I^2/K^2$, these limits correspond to a scatter of U-values from $17^2/100 = 2.9$ to $23^2/100 = 5.3$, an average of about 1.2 units about the theoretical value $U = 4$.

For a much higher value of U, say 36, $I = 10\sqrt{36} = 60$. The average limits of I will again be $\pm$ 3, from 57 to 63. However, in the $U = I^2/K^2$ formulation, these limits of I correspond to average limits of U from 32 to 40, a scatter of 4 units about the value 36. The scatter of U is about three times as large when $I = 60$ as when $I = 20$.

Therefore, $I = K\sqrt{U}$ is the simpler formulation to use because no matter where we are on the line, its average scatter of $(I - K\sqrt{U})$ can be denoted by $\pm$ 3. (If needed, the size of the scatter of U values can still be estimated with this formulation, as just demonstrated.)

Exercise 6E. Prediction and Decision-making

An equation $y = ax + b$ has been found to fit previous data within limits of $\pm c$. The prediction of y in future data will be that y is distributed about $(ax + b)$ within average limits of c. Is this of any use for decision-making, where a single value of y is usually needed?

Discussion.

Prediction and decision-making are separate processes. For example, suppose that we have to decide on the maximum load *L* which a certain bridge can take when it is made of girders of a certain thickness *T*.

First we have to arrive at a relationship $L = aT + b \pm c$ between maximum safe loads *L* and thickness of girders *T*. From this we predict that *L* is distributed about $aT + b$ within average limits $\pm c$.

Then we have to make a decision about the maximum load to be allowed. One possibility is to choose the average value $(aT + b)$, but this is unlikely

when dealing with *safety*. Another might be to use the engineer's adage of "multiply by three for safety" and hence use $(aT+b)/3$ as a maximum permissible load. Another might be to restrict the load by some multiple of the average scatter c, e.g. $(aT+b)-5c$, being about the highest load-level at which no bridge with girders T has ever collapsed, but this restriction might be too costly.

The decision here clearly involves considerations of risks and costs. These are separate from the task of predicting at about what load the bridge would actually collapse.

CHAPTER 7

Deriving a New Relationship

In this chapter we discuss fitting a linear equation to measurements of two variables. The situation considered is where there are two or more sets of readings. The ,probIem of having to fit such an equation arises primarily with data that are being handled for the first time.

7.1 The First-time Problem

Treating data as if they were being analysed for the first time is fairly common, but strictly speaking such a first-time situation can happen only once. After that, there must be a previous 'result: e.g. that $y \doteqdot 5x + 10$, with which the new data can be compared, as discussed in Chapter 6.

At times one may not know about any earlier results, or they may not be in a usable form. Sometimes it is easier to fit a first-time. equation to the new data and compare it with the prior results. Or the previous equation may not fit and we want to find another equation to summarise the new data. These are proper reasons for having to fit an equation.

In other cases we should use the previous results in the analysis. But people are often afraid to use earlier findings because the conditions were different (the earlier results came from Mexico, or were pre-war, or whatever) and are thought not to be comparable. This view is wrong One cannot know whether or not the relationships in different sets of data are the same without comparing them. In general, the aim is to study a relationship under a variety of different conditions. If the same result emerges, it is the more powerful just because the conditions were so different. If the results are different, that also is highly informative: the relationship was different last year, or in Mexico, etc.

7.2 A Degree of Prior Knowledge

To illustrate the fitting of a new relationship we again use the data on the heights and weights of children. But we suppose now that we do not know of

any relationship fitted to previous data. Nonetheless, a good deal of other background knowledge already exists.

For example, everyone knows something about children's heights; that children grow taller as they get older, boys tend to be taller than girls, some races are shorter than others, tall parents tend to have tall children. We know that height does not vary hourly or daily, that individual children vary a good deal from each other (some girls are taller than some boys), and that despite this there are statistical regularities, so that boys *on average* are taller than girls. We also know that age is generally the dominant factor in accounting for children's height. Any comparison of the heights of different children (boys with girls for example) has to be done *at the same age* if it is to be meaningful.

There is similar knowledge about children's weights. In addition, weight can vary marginally in the short-term (e.g. by eating a 2 lb meal) and sometimes fairly markedly in the medium-term (people losing or gaining weight).

Something is also known about the relationship between the two variables. The two measures do not always go together, taller children usually weigh more than shorter ones, but there are also tall "thin" children and short "fat" ones. We also know that height and weight both increase as children get older: there is no doubt that as children grow, height and weight are correlated.

But at this stage we do not know the quantitative form of the relationship. Nor do we have much theoretical insight into it, or know which other factors (such as sex, race, age, country, etc.) influence it. These are the questions that we have to start answering in our "first-time" analysis.

7.3 **The Design of the Study**

Our background knowledge helps in choosing what kind of data to analyse. These may be selected from available sources or be newly collected. With children's heights and weights many measurements already exist. Simple findings can be established if a purposeful approach to the available data is adopted.

One obvious aim in any study is to select children who differ from each other in their heights and weights, so that there is some variation to analyse. **Since we** know that certain races tend to be taller than others (e.g. Caucasian versus Chinese), we can study how racial differences in height relate to those in weight. But since we know that age is the largest factor affecting children's heights and weights, an efficient design for a first study is a comparison of children of different age-groups.

Following the growth pattern of a group of children over a period of years (a so-called "longitudinal" study) is, however, expensive and technically difficult. We can therefore either turn to another factor to study first, or

analyse children of different ages measured at the same point-in-time (a "cross-sectional" study). This is very simple. Most available height and weight data for children are already grouped by yearly ages (or in school-classes of children of *similar* ages). An example for three age-groups is given in Table 7.1.

TABLE 7.1 The Average Heights and Weights of Ghanaian Privileged Boys Aged 11, 12 and 13

Ghanaian Boys	Age (in years)		
	11	12	13
Average height (ins.)	57	58	59
Average weight (lbs.)	83	86	88

It is accidental that age, the main design variable used here, is itself quantitative. This will not affect anything we do in the analysis during this chapter, In principle, the data could equally well consist of groupings of children who differ *qualitatively, e.g.* by race, sex, religion, school-grade, socio-economic condition, colour of eyes, and so on.

Having used our prior knowledge to control the data selection, we are now not faced with an unstructured set of height and weight readings, but with data which are ordered into several sub-groups. And as expected, the older children are both taller and heavier. All we have to do is describe the relationship.

Each reading in Table 7.1 is the average of a group of about 50 children of the stated age. This produces more regular statistical patterns, since individual children generally differ from each other considerably. In other kinds of studies one may have to make do with a single reading in each sub-group (as with the time-series data analysed in Chapters 1 and 2). This can affect the scatter or precision of the results, but not the principles of the analysis.

7.4 A First Working-solution

A glance at Table 7.1 shows that average height increases by an inch per year, and weight by 2 or 3 lbs. The increases in height and weight are therefore in roughly similar ratios from year to year, so that the relationship is more or less linear, as is also shown by plotting a graph like Figure 7.1. As a simplifying approximation it is therefore worth fitting a linear equation of the form

$$w = ah + b$$

to summarise the data, rather than look for some kind of curve at this stage.

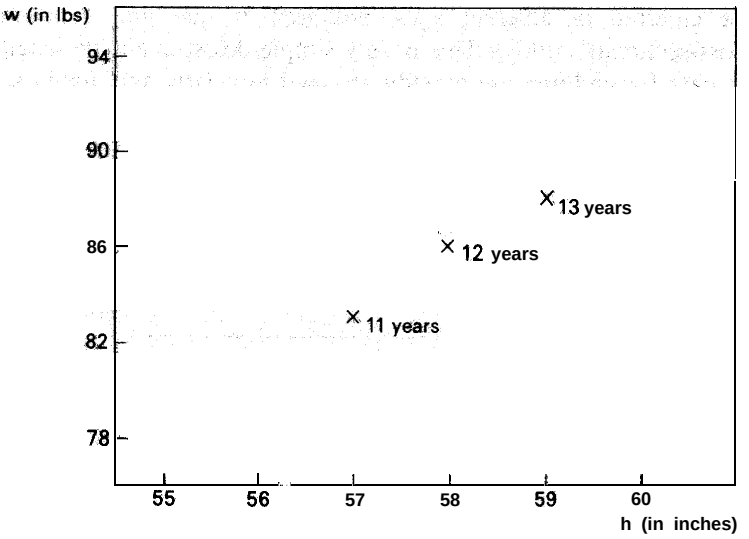

Figure 7.1 The Average Heights and Weights of 11-, 12- and 13-Year-Old Ghanaian Boys

But since the three points do not lie *exactly* on a straight line, any linear equation will be a deliberate oversimplification to provide a simple, approximate summary of the data. No one straight line will reflect the data perfectly, and there can be no unique answer. The equation that we choose will therefore be no more than a tentative initial working-solution.

One way of determining a value for the slope-coefficient a in such a working-solution is by the formula

$$a = \frac{y_2 - y_1}{x_2 - x_1},$$

where (x_1, y_1) and (x_2, y_2) are the two extreme pairs of mean values in the data. For the data in Table 7.1, they are the readings for the 11- and 13-year-olds: i.e. 57 inches and 83 lbs, and 59 inches and 88 lbs. This gives

$$a = \frac{88 - 83}{59 - 57} = \frac{5}{2} = 2.50,$$

working to three significant figures at this stage. So far the equation therefore reads

$$w = 2.50h + b.$$

To determine a value of the coefficient *b*, *we* put the line through the overall averages of the readings, which are 58.0 inches and 85.7 lbs. This gives

$$85.7 = 2.50 \times 58.0 + b,$$

$$b = 85.7 - (2.50 \times 58.0) = -59.3.$$

The resulting equation is

$$w = 2.50h - 59.3.$$

For a given value of *h*, our theoretical estimate of w is therefore $(2.50h - 59.3)$, as shown in Table 7.2.

TABLE 7.2 The Theoretical Estimates (2.50h - 59.3)

Ghanaian Boys	Age 11	12	13	Av.
Av. height: h	57.0	58.0	59.0	58.0
Av. weight: w	83.0	86.0	88.0	85.7
2.50h - 59.3	83.2	85.7	88.2	85.7

The differences between our theoretical estimates and the observed readings of w show how well the initial linear solution fits the data and are given in Table 7.2a. (With a range of only 5 lbs in the average weights it is helpful to calculate the theoretical values to one place of decimals. Hence the observed data are also shown to one place in the table.)

TABLE 7.2a The Deviations from the Working-Solution w = 2.50h - 59.3

Ghanaian Boys	Age 11	12	13	Av.
Deviations: w - (2.50h - 59.3)	-.2	.3	-.2	.0*

* Average size ignoring sign = 0.2 lbs

The important feature of these deviations is that they are irregular in sign, + − +. Their average size ignoring the sign, called the *mean deviation*, is about 0.2 lb. This is small compared to the 5 lb total range of weights in the data, so the equation provides a fairly good fit.

The rationale behind this method of deriving an initial working-solution is that any reasonable equation must go more or less through the overall means of the data, and that fitting the slope by the extreme values will tend to leave irregular deviations, if the data are in fact more or less linear. As a result the data are easy to summarise along the lines discussed earlier,

namely that w tends to vary as $(2.50h - 59.3)$, with deviations which are irregular and of a certain average size (here 0.2 lb).

Sometimes the extreme values used to determine the slope-coefficient are not as clear as in this case. Figures 7.2A and B give two examples. In such situations some more *ad hoc* approach can be adopted to derive a first working-solution, like excluding an odd value or grouping several extreme readings together to determine the slope-coefficient. Of necessity the fit will not be close and the solution will be particularly tentative.

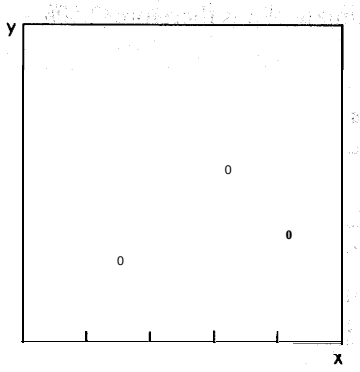

Figure 7.2A Which extreme Values?

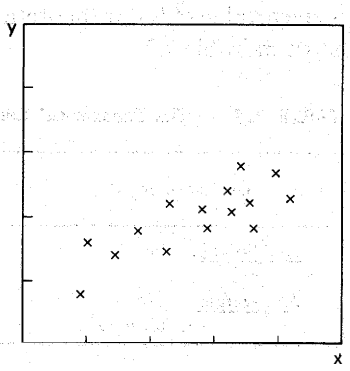

Figure 7.2B Another Example

7.5 **The Existence of Alternative Solutions**

Since the initial equation above did not fit the data perfectly, there will be other equations which might be just as good or even better. The differences in the fit of such alternative working-solutions are usually slight. It is therefore difficult to choose one as being clearly the best. But this also means that descriptively it is not important which equation one uses.

One alternative is the equation $w = 3h - 88.3$, which is obtained by trying a slope of-3 and again putting the line through the overall means of the data. "Trying a slope of 3" may not seem a very rigorous method of analysis, but the method of derivation matters far less than the results. The two equations in fact give very similar theoretical values of w if we insert the values of *h* for the 11-, 12- and 13-year-old boys:

	h	=	57	58	59
Original:	$2.5h - 59.3$	=	83.2	85.7	88.2
New:	$3.0h - 88.3$	=	82.7	85.7	88.7

The new equation should therefore also fit the observed data fairly well, and Table 7.3 shows that it does. There is one sizable deviation of -0.7 lb,

TABLE *7.3* *The* **Fit of the Alternative Working-Solution w =** 3. Oh - 88.3

Ghanaian Boys	Age 11	12	13	Av.
Av. height: h	57.0	58.0	59.0	58.0
Av. weight: w	83.0	86.0	88.0	85.7
3.0h - 88.3	82.7	85.7	88.7	85.7
w - 3.0h + 88.3	.3	.3	-.7	0.0*

* Average size ignoring sign = 0.4

and the mean deviation is 0.4 lbs. Compared with the 5 lb range of weight, this is only fractionally bigger than the mean deviation of 0.2 lbs for the original equation. The two equations look fairly different, however, both in their slope-coefficients and their intercept-constants. Nonetheless, they give very similar results. Indeed, they differ hardly more from each other than either does from the data. But this is only over a limited range of variation. Outside. this range the equations will differ increasingly from each other.

Figure 7.3 illustrates the point with four different equations: the two straight lines just discussed, and two curves, A and B. The four equations fit

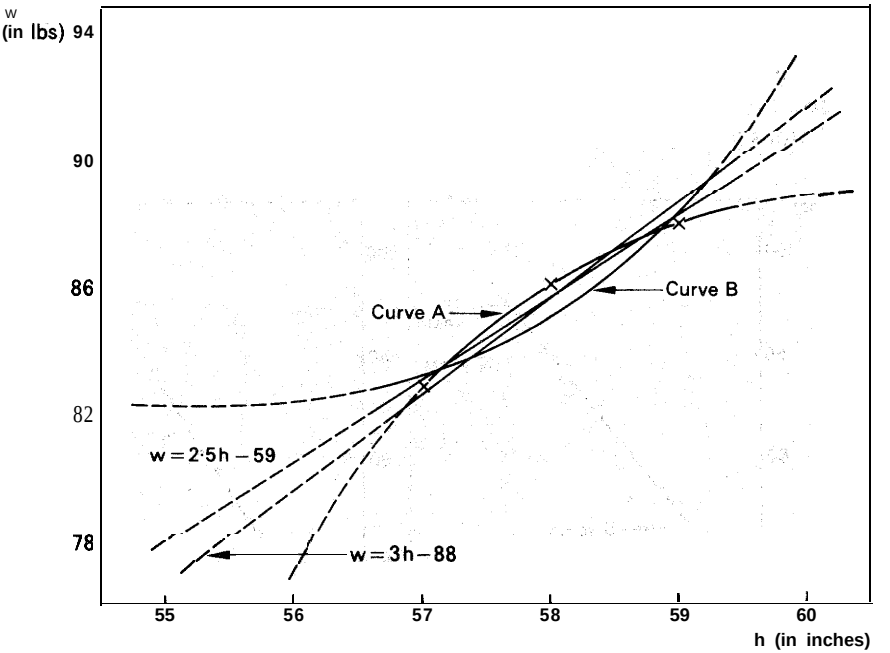

Figure 7.3 Alternative Working-Solutions for the 11- to 13-Year-Olds

the given data almost equally well, but differ markedly from each other *outside* the range covered by the 11- to 13-year olds. It follows that the apparently arbitrary choice of a working-solution for the given data can be narrowed by seeing which equation, if any, also holds for boys outside the initial range of variation,

We therefore now turn to a wider range of data, namely for the 5- to 13-year old privileged Ghanaian boys referred to in the previous chapter. Table 7.4 sets out the data. The weights here vary by 40 lbs, so that we can drop the third significant figure in the coefficients and generally work to the nearest whole pound, keeping an extra decimal place only in the overall averages for working purposes.

TABLE 7.4 The Failure of the 11- to 13-Year Olds' Equation $w = 2.5h - 59$ to Fit for Younger Boys

Ghanaian Boys	Age									Av.
	5	6	7	8	9	10	11	12	13	
Av. height: h	45	48	49	51	54	55	57	58	59	52.9
Av. weight: w	46	52	58	62	70	75	83	86	88	68.9
$2.5h - 59$	54	61	63	69	76	78	83	86	88	73.1
$w - 2.5h + 59$	-8	-9	-5	-7	-6	-3	0	0	0	-4.2*

*Average size (5-10 years) ignoring sign = 6 lbs

Table 7.4 shows immediately that our initial working-solution $w = 2.5h - 59$ does not fit the younger boys at all. It gives large and consistently negative deviations for boys under 11 years. Figure 7.4A also illustrates this. It follows

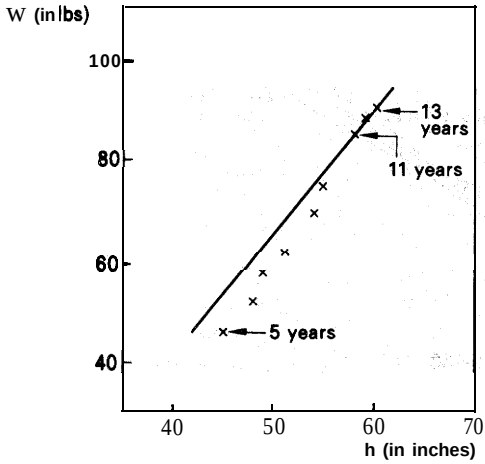

Figure 7.4A The Failure of $w = 2.5h - 59$

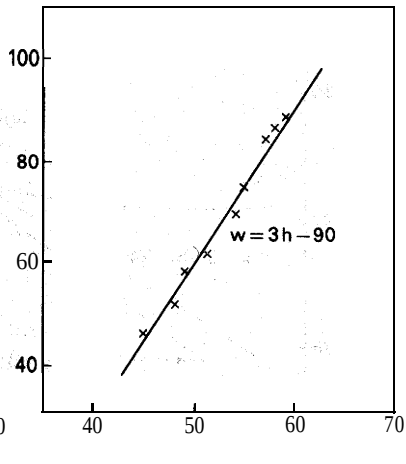

Figure 7.4B A New Working-Solution

that our earlier worry about precisely which equation to fit to the 11- to 13-year olds has been superseded by a much bigger problem.

However, the new situation is not hopeless: a strong relationship between height and weight clearly exists and can no doubt be described by an adjusted linear solution with different coefficients.

7.6 A New Working-solution

To find an adjusted linear working-solution we can use the same procedure as before, but this time we apply it to the wider range of data in Table 7.4.

The slope-coefficient is again determined from the two extreme pairs of readings, i.e. we divide the range of the weights by the range of the heights to get $a = (88 - 46)/(59 - 45) = 3.0$. The intercept-coefficient is calculated by requiring that $w = 3h - b$ should hold for the overall averages of all the data, so that $b = 68.9 - (3 \times 52.9) = -89.8$, or -90 to two significant figures. The new working-solution is therefore

$$w = 3h - 90.$$

Table 7.5 shows the fit of this equation. The individual deviations appear fairly irregular and their average size is 1.3 lbs, which is small compared with the 42 lbs range of average weight in the data.

TABLE 7.5 The New Working-Solution w = 3h - 90 Fitted to the Data for the 5- to 13-Year-Olds

Ghanaian Boys	Age 5	6	7	8	9	10	11	12	13	Av.
Av. height: h	45	48	49	51	54	55	57	58	59	52.9
Av. weight: w	46	52	58	62	70	75	83	86	88	68.9
3h - 90	45	54	57	63	72	75	81	84	87	68.7
w - 3h + 90	1	-2	1	-1	-2	0	2	2	1	.2*

* Average size ignoring sign = 1.3 lbs

The new equation, $w = 3h - 90$, therefore gives a good fit to the data as a whole. But it does not fit the 11- to 13-year-old boys as well as the earlier working-solution did. The deviations for these ages are now 1 or 2 lbs, compared with only about 0.2 lb 'before. Furthermore, the deviations are all positive, all three points lying *above* the fitted line. The choice facing us is between an equation which gives a close fit to a limited range of data and one which covers a much wider range of readings but less precisely, as

Figures 7.4A and B illustrate. The answer is that it is descriptively and conceptually much simpler to deal with one equation plus some irregular scatter for all the data, rather than with various *different* equations with smaller scatter for different parts of the data. This determines the approach described here.

7.7 **Alternative Working-solutions**

On closer scrutiny the deviations in Table 7.5 seem to have a slight systematic + − + pattern. They are positive for two of the three youngest age-groups, negative or zero for the middle three age-groups, and positive again for the three oldest age-groups. This may seem like reading too much into just nine readings, but at this early stage in the analysis we are only noting possibilities. A *curved* working-solution might therefore fit better in giving less of a pattern, especially for the 11- to 13-year olds.

This analysis brings us back again to the problem of choosing among different working-solutions. For example, curve C in Figure 7.5A gives less regular deviations than the working-solution $w = 3h - 90$. Almost the same effect can, however, also be achieved with another *linear* working-solution, such as

$$w = 3.2h - 100,$$

shown in Figure 7.5B. The equation was derived by applying the usual fitting-procedure to *groupings* of the more extreme average heights and weights, namely the 5- to 7-year olds and the 11- to 13-year olds. This reduces

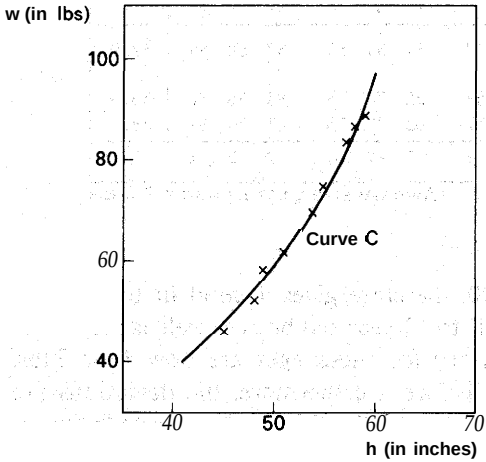

Figure 7.5A A Possible Curve

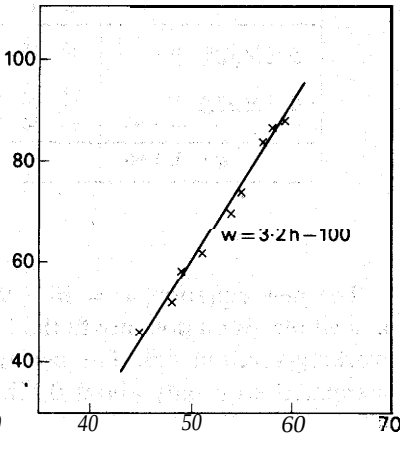

Figure 7.5B An Adjusted Line

the reliance on some particular extreme value which may be a little abnormal. (Figure 7.4B for example suggests that the reading for the 5-year olds influenced the earlier equation too much. Without it we would have fitted a steeper line, more like the new equation $w = 3.2h - 100$.)

The mean deviations of the two new alternative equations in Figures 7.5A and B are about 1.2 or 1.3lbs, which is virtually the same as for the ear-her equation $w = 3h - 90$ in Figure 7.4B. The important difference is that the deviations have a less regular pattern, in particular those for the three oldest age-groups.

There is therefore once more a certain variety of equations which fit the data "reasonably" well. We still have a problem of choice. But the degree of uncertainty has been greatly decreased by extending the range of variation covered, Figure 7.6 shows that three of the earliest working-solutions from Figure 7.3, $w = 2.5h - 59$ and Curves A and B, are no longer tenable at all, and that the fourth, $w = 3h - 88$, is also ruled out by its small but consistent bias.

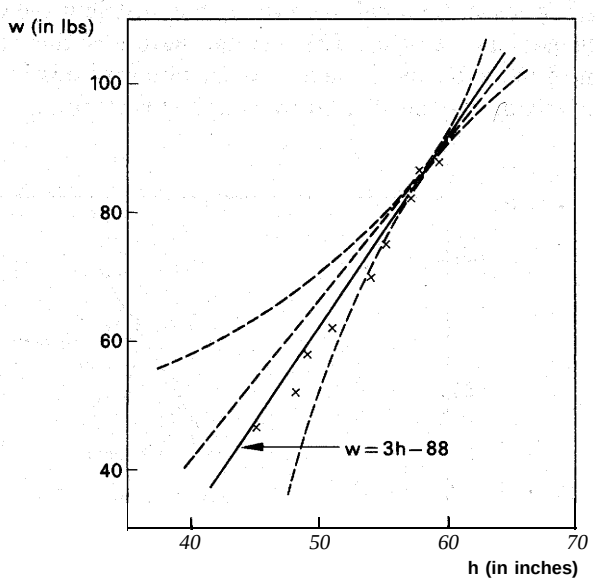

Figure 7.6 The Initial Working-Solutions of Figure 7.3

The ambiguity which now remains will be reduced by analysing yet more data, coupled in due course with the growth of theoretical understanding. The question is which model, if any, will describe not only the heights and weights of these particular 5- to 13-year-old Ghanaian boys, but other data

sets as well? The analysis of further data will be discussed in the next chapter, together with the first stages of dealing with *curved* relationships.

But at this stage, with only nine pairs of readings, there is no need to worry unduly about precisely which of the possible equations to select. If the further data show quite a different height/weight relationship (with all the new data for example lying well to the left of the line in Figure 7.5B), then it no longer matters which particular equation was fitted. And if the new data more or less agree with the present readings, they will narrow the choice of equations.

7.8 **The Scatter of Individual Readings**

The analysis of a relationship essentially has two parts. One describes the systematic variation of the mean values between different groups of readings, as we have been doing. The other describes the scatter of the *individual* readings in each group, if there is more than one reading in each. This is a separate part of the analysis. It generally does not affect the question of which equation to fit.

Figure 7.7 gives a notional picture of the distribution of individual children's heights and weights. The size and nature of this scatter cannot affect the line fitted to the mean values (except with very small samples, when the observed means may be affected by sampling variation).

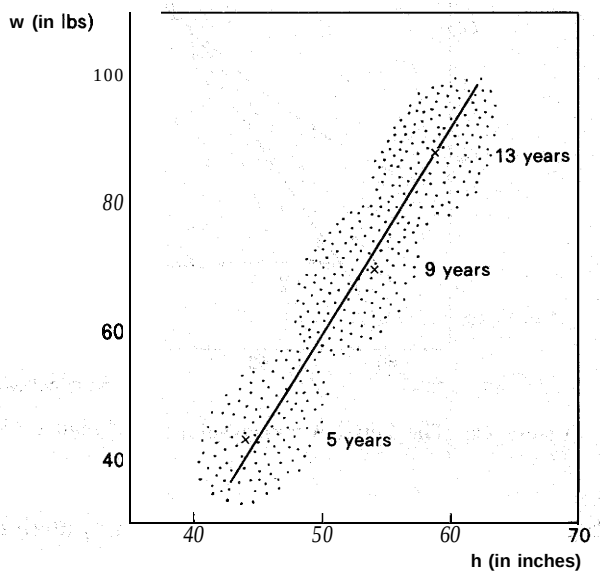

Figure 7.7 The Distribution of Readings for Individual Boys (A notional representation)

We therefore first fit an equation like $w = 3.2h - 100$ (together with any discrepancies in approximating to non-linear data). Then we describe the scatter of the individual readings by summarising the average size of the deviations ($w - 3.2 + 100$) for individual boys.

From the information on Birmingham boys supplied privately by Dr. E. M. B. Clements (1954), as mentioned earlier, we can estimate the average size of such individual deviations to be about 8 lbs, as a broad average over all age-groups. This provides a better "feel" for the data, and is necessary knowledge for certain practical applications of the results (as was illustrated in Section 6.4).

However, such information is often not reported. For example, none of the published sources of the height and weight data referred to earlier (see Table 5.1) gave information on the scatter of the individual readings. As mentioned in Section 5.2, such an omission is open to purist criticism. But, in practice, people are mostly concerned with the *systematic* variation in their data, how w varies with h in different sets of data, and this is represented by the equation fitted to the age-group means.

7.9 Summary

In this chapter we have discussed fitting a linear equation to two variables. This occurs mainly when there is no previous information about the relationship. But even when handling data for the first time, one usually has some background information about the variables in question. For example, it may already be known where the larger variations occur. One can therefore choose to analyse groups of data that differ from each other systematically.

Even for data which are more or less linear, the means of different groups of readings usually do not lie *exactly* on a straight line. Various alternative approximations are therefore possible in fitting a straight line. An initial linear working-solution can be obtained by calculating the slope-coefficient from the ratio of the differences between the highest and lowest mean values of the two variables, and the intercept-coefficient by making the line go through the overall averages. This procedure generally leads to residual deviations which are irregular and hence easy to summarise simply in terms of their average size.

Because the different sets of data do not lie exactly on a straight line, no linear equation can fit perfectly. The equation fitted is therefore one of various possible working-solutions_ These all give a similar fit to the data and therefore all tell effectively the same story. In general, two solutions which fit the data with irregular deviations will differ no more from each other than each does from the observed data.

Initial working-solutions are adjusted in subsequent work, mainly to accommodate additional data. This occurs especially if the new data lie

outside the initial range of variation. The degree of uncertainty involved in choosing among different possible equations is therefore reduced. The aim of the analysis lies not so much in finding an equation that fits best for the initial set of readings, but one that can generalise to a wider range of data.

CHAPTER 7 EXERCISES

Exercise 7A. Problems with Extreme Values

Fit an initial working-solution to the following five pairs of readings:

x:	15	10	20	18	12
y:	7	3	8	11	1

Discussion.

This example typifies a fairly common problem when fitting an equation by the method in Section 7.4, namely that it is difficult to determine the extreme values in the data. Ordering the readings by the size of the x variables gives

x:	10	12	15	18	20
y:	3	1	7	11	8

Using the two extremes we get a slope-coefficient of $(8-3)/(20-10)=.5$ and the equation

$$y = .5x - 1.5.$$

But if we order the reading by the y variable, we have

x:	12	10	15	20	18
y:	1	3	7	8	11

Using the two extremes here gives a slope-coefficient of $(11-1)/(18-12)=1.7$ and the equation

$$y = 1.7x - 19.5.$$

These two equations look very different.

The problem arises because although there is a general tendency in the data for high x to go with high y and low x with low y, this is not the case for the two *lowest* or *highest* pairs of values. The lowest x-value is not the lowest y-value, even though both are *low*.

In a case like this a better alternative working-solution can be fitted by combining the two lowest pairs and the two highest pairs of values, as suggested in Section 7.4. Fitting a slope-coefficient to these grouped means gives $(9.5-2)/(19-11) \doteqdot 1$ and the equation

$$y = x - 9 .$$

This is clearly a compromise between the first two equations.

The reader can check that each equation fits the y-values with a mean deviation of about 2 units. There is therefore not a great deal of difference among them in their average fit. But the "compromise" solution, $y = x - 9$, is more attractive because it depends less on an isolated and rather erratic

extreme value and because it has the more evenly sized and irregular deviations.

The main conclusion is that if the scatter in the data is relatively large, then there will be a considerable range of equations that can reasonably be fitted. Additional data will generally reduce the ambiguity, as already stressed in the main text of this chapter.

Exercise 7B. Subjective or Objective?

Are the analysis procedures described here very subjective'?

Discussion.

No. If the analyst describes what he has done, any experienced person will arrive at effectively the same result by applying the same procedure to the same data. This is the criterion of objectivity in analysis. Wat is more, if the result generalizes, anyone will be able to arrive at it by following the same procedure with *other* data. This is the criterion of objectivity in *science.*

Objectivity should not be confused with the absence of choice. For a method to be "objective" does not mean that any fool must always get the same answer as a highly experienced analyst. It is not necessary for there to be only one possible method or one possible result.' A foolproof method of being objective would be always to fit the equation $y = 2x + 3$. Everyone would then very objectively get the same answer for any data but generally be wrong.

Exercise 7C. Alternative Working-solutions

How can one claim that two alternative "working-solutions" like

$$w = 2.5h - 59.3$$

$$\mathbf{w} = 3.0h - 88.3$$

in Section 7.5 say more or less the same thing about the data? According to one, w varies 20% more with *h* than according to the other. And when *h* is 0, one says that w is 59 lbs and the other 88 lbs. Such differences are not negligible.

Discussion.

The two equations do not purport to say at precisely what rate w varies with *h*, since this was not altogether clear from the given three pairs of readings. That is why different working-solutions are possible. Nor do such equations purport to reflect the values of w for values like $h = 0$ which are way outside the observed range.

Instead, the equations merely aim to summarise the observed sets of data. Within the observed range of variation they generally differ from each other no more than each differs from the data. (The example in Exercise 7A illustrated an extreme type of case where some working-solutions can differ markedly.)

Once data covering a more extensive range of variation have been successfully analysed, the coefficients of the equation may bear more direct interpretation. Thus for the 5- to 13-year olds analysed in Section 7.6, it would appear that weight varies by something like 3.0 to 3.2 lbs for

every inch increase in height, and working-solutions like $w = 3h - 90$ and $w = 3.2h - 100$ summarise this.

But the "intercept-coefficients" of -90 or − 100 still have no direct meaning. They do not say what w would be for $h = 0$, since no such data have been observed.

Exercise 7D. Why the Problem?

Why is there this problem of different possible working-solutions?

Discussion.

Because one is forcing a linear equation on to non-linear data. Approximation is often said to be an art, and within the small limits of choice illustrated here this is to some extent true.

Exercise 7E. The Intercept-coefficient

Why do the equations fitted to the heights and weights not go through the point (0, 0), since initially children have virtually no height or weight?

Discussion.

The data being analysed do not cover babies, let alone pre-natal conditions.

Some attention can, however, be paid to such external knowledge: for example to differentiate between two otherwise equally possible working-solutions. But it would be wrong to force the empirical equation too much, since at this stage we do not know what mathematical form the height/weight relationship should take below 5 years.

Exercise 7F. The Wrong Working-solution

Suppose that the data in Exercise 7A have been summarised by the working-solution $y = x - 9 \pm 2$, in the range of $x = 10$ to $x = 20$. How does this assist in analysing the following additional data?

x:	12	14	27	30	42
y:	3	6	9	14	18

Discussion.

The earlier result does not fit the higher values here. For example, for $x = 42$, the predicted value is $42 - 9 = 33$, instead of the observed value 18.

Fitting a new working-solution to the new data gives $y = .5x - 2.5$. We must now decide whether this will also fit the previous data, without having direct recourse, to them. All we know about the earlier data is that they were fitted by $y = x - 9 \pm 2$, within the range of $x = 10$ to $x = 20$.

We therefore compare the two equations in this range:

x	10	15	20
.5x − 2.5	2.5	5.0	7.5
x - 9	1.0	6.0	11.0
Difference	1.5	− 1.0	-3.5

The average difference is about 2, which is the same as the reported fit of the earlier equation to the data. The *maximum* discrepancy is -3.5, which is probably no larger than the largest discrepancy for the earlier data. This implies that the equation $y = .5x - 2.5$ will fit the original data roughly as well as the first working-solution, $y = x - 9$, did.

This conclusion has been reached merely from a *summary* of the earlier data, without having seen the detailed readings. The initial working-solution proved "wrong" in the light of new data, but it still performed its primary function of first adequately summarising the original data and then leading to a better working-solution for both that *and* the new data.

Exercise 7G. **Different Sets of Data**

The analysis in this chapter depends on having more than one set of data, and these must have different means. (Otherwise no equation can be fitted by the procedure that has been described.)

What happens if we either have only one set of readings, or if our different sets of readings differ little (if at all) in their means?

Discussion.

Having only one set of readings cannot happen often, since any worthwhile study has to be repeated. If all the different sets of data have the same means, the observer has not managed to exercise any effective control over his variables: there may be variation in x and *y*, and this may correlate, but it is effectively uncontrolled or "error', variation. The observational basis for establishing an empirical generalisation does not exist.

The observer needs to see if he can differentiate his data more effectively by using better selection criteria (e.g. older boys versus younger boys rather than size of family, say).

A set of height and weight readings for a group of *n* children consists of a *single* set of readings if we know nothing about each boy other than his height and weight. But if for each separate child we have a description of its age, sex, race, place of residence, number of siblings, etc., and if the children differ from each other in some or all of these respects, then we have *n* different sets of data, each consisting of a single child. (Much observational data consists of 1 reading per set.)

Exercise 7H. Height as a Function of Weight

In Section 7.6 we fitted the working-solution $w = 3h - 90$, which expresses weight in terms of height for the 5- to 13-year-old Ghanaian boys (Table 7.5). What would we get if we used the same technique to fit an equation of the form $h = pw + q$, expressing height in terms of weight?

Discussion.

From the data in Table 7.5, the slope-coefficient in the equation $h = pw + q$ will be $(59 - 45)/(88 - 46) = 14/42 = .33$. The intercept-coefficient will be $52.9 - .33 \times 68.9 = 30$. This gives the equation

$$h = .33w + 30.$$

Dividing through by .33 (or multiplying by 3) gives

$$3h = w + 90,$$

which is $w = 3h - 90$, as before.

As long as the same method of fitting and the same "extreme points" are used, the same equation is obtained whichever way round it is written.

Exercise 7I. Height and Age, and Weight and Age

In Table 7.5 age is a third quantitative variable. What are the relationships between height and age and between weight and age?

Discussion.

Fitting a working-solution to the data in Table 7.5 for height h and age A gives $h = 1.75A + 37$. This has a mean deviation of about $\pm .5$ inches, and the deviations appear irregular.

For weight and age, the working-solution is $w = 5.25A + 22$, with a mean deviation of $\pm$ 1.6 lbs. But here the deviations show quite a marked pattern; negative for low ages, positive for 9 to 12 years, and negative for 13 years. This suggests that a curve might fit better than a straight line, as is discussed in Exercise 8F.

Exercise 7J. The Velocity of a Falling Body

The velocity of a falling body was measured at intervals of one second with the following results

Time t (in sec):	0	1	2	3	4	5
Velocity v (in ft/sec):	10	45	70	120	130	160

What relationship would you fit?

Discussion.

It is well-known in physics that a body falling freely near the surface of the earth is subject to a virtually constant acceleration (an increase in its velocity) of about 32 feet per second every second, i.e. v varies as $32t$. Given that the observed velocity at time $t = 0$ was 10 ft/sec, the velocity at any time t should therefore be

$$v = 10 + 32t.$$

This theoretical result gives a good fit to the data. It is a typical case where a "first-time" method of fitting an equation is unnecessary.

Exercise 7K. Fitting to Individual Readings or to Means

Why are working-solutions fitted to the mean values of the different sets of data and not to the individual readings of x and y?

Discussion.

Suppose there are two or more sets of data, such as the different age-groups in our numerical example. Any line fitted to the first set has to go through its mean values. If the same line is to hold for the second set of data, it also has to go through the means of *that* set. These two points determine the line. The individual readings **do not** affect the line that is fitted.

Exercise 7L. The Analysis of the Individual Readings

How can one analyse the scatter of the individual readings in each set of data? For example, in Figure 7.8A would it be useful to fit a separate equation to the individual readings in each age-group?

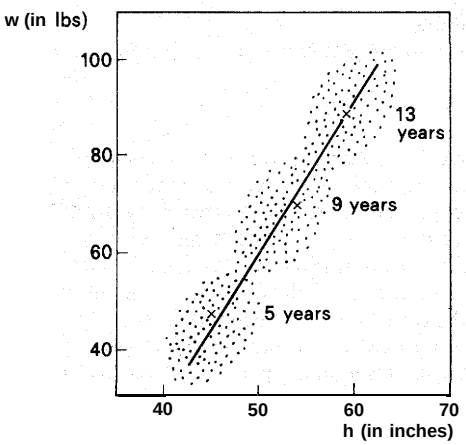

Figure 7.8A Individual Boys (Figure 7.7)

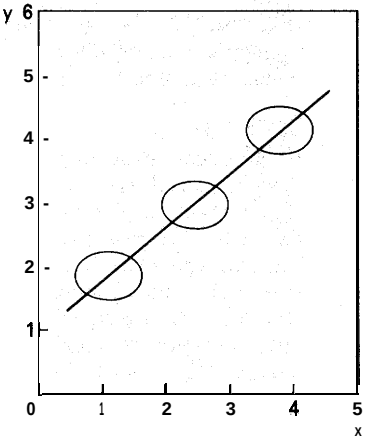

Figure 7.8B A Different Form of Scatter

Discussion.

If equations are fitted to the separate age-groups in Figure 7.8A which differ from the equation already fitted to the group means, they will also differ from each other. We would therefore generally finish up with n unique equations for n sets of data. It is not clear what use they could be.

However, the data in each age-group might not be "in line" with the fitted equation, the scatter of the individual readings taking a different shape, as illustrated in Figure 7.8B. This would merit further scrutiny. One possible cause for data to be in this form is simply if the errors of measurement of one variable (here x) are much greater than those of the other.

The scatter of individual readings can be broken down into two types of "error", one attached to each variable:

$$x = \text{"true value of x"} + \text{"error of } x\text{"},$$

$$y = a(\text{"true value of x"}) + b + \text{"error of } y\text{"}.$$

But even though in the past I have engaged in some theoretical research into this kind of model (e.g. Ehrenberg, 1950, 1951), I have found virtually no call for it in practical data analysis.

In terms of our example, the main practical questions in a more detailed analysis of individual deviations seem to be (a) whether the individual boy's deviations from the relationship are consistent over time, i.e. does a boy remain overweight if he was overweight earlier? (b) how these deviations relate to other factors, such as the height/weight patterns of parents and of siblings the boy's medical and nutritional history, and so on. To answer

such questions more information is needed than just the boys' heights, weights, and ages. In particular, the *longitudinal* type of design mentioned in Section 7.3 would be required, giving repeated measurements of the same boys over a period of time.

Exercise 7M. Is There a Relationship?

How can we tell whether a relationship exists between two variables for some given data?

Discussion.

An equation of the form $y = ax + b$ reflects a numerical association between the observed values of x and y. A relationship exists if the average deviation of y from the line, $(y - ax - b)$, is less than the average deviation of the y-values from their overall mean $\bar{y}$, $(y - \bar{y})$. If the difference between the two types of deviation is small, the relationship is weak. (If the data are based on random samples and the number of readings is small, one must test the statistical significance of the result, as is discussed in Chapter 18: an association between x and y could exist in the sample due to chance errors in the sampling, without an association in the population sampled.)

This question of whether the two variables are related can only occur at the "first-time" stage of analysis. After that there will be prior information about what has already been found in other data. The more common question in practice is therefore not whether the relationship exists, but whether the relationship in one case is the same as in others.

Lack of correlation should not in any case be regarded as a *failure.* For example, establishing an empirical generalisation consists of showing that one's results (e.g. the values of the coefficients in the equation $w = 3h - 100$) are *not* related to other variables.

In general, nothing could be simpler, and hence more important, than to show that y does *not* vary with x. This is especially important if the analyst firmly expected that y *would* vary with x. Having waited a minute or two to overcome his discomfiture, he should be able to say "So y is *not* related to x-what can I make of that?".

Exercise 7N. Other Methods of Analysis

By reference to statistical textbooks and journals, discuss other methods of fitting a straight-line equation.

Discussion.

Several different approaches have been considered in the statistical literature. The main ones are as follows.

Regression Analysis. Here the coefficients of the equation $y = ax + b$ are determined by making the sum of all the squared deviations $(y - ax - b)^2$ as small as possible. This is discussed in Chapter 14.

Regression analysis is generally applied to a single set of readings, which is a different situation from that considered in the present chapter. The problem of fitting a regression equation to two or more distinct sets of data does not appear to be discussed in the literature. Nor is it usually claimed that regression analysis leads to generalisable results.

Regression Applied to the Group Means. A regression equation could be fitted to the mean values of different sets of readings, in effect treating them as a single set of data. But if the means lie exactly on a straight line, there is no fitting problem. If they do not, one would be fitting a linear regression equation to non-linear data, therefore the statistical requirements and advantages of regression theory would not apply. Thus regression analysis applied to the group means has no particularly attractive properties.

Regression Applied to Pooled Data. A regression equation could also be fitted to two or more sets of data by first pooling all the readings. But this would lose all the information about how the readings differ from each other (e.g. boys of different ages). The regression equation would also depend on the arbitrary numbers of readings in the different groups. Furthermore, pooling is unnecessary because an equation can be fitted to the group means, as discussed.

Functional Analysis. This method is based on the "errors in both variables" model outlined in Exercise 7L. But as described in the statistical literature functional analysis is generally discussed in terms of a single set of data. In such a case it is agreed that the method cannot actually provide a solution, as the slope-coefficient a in this model cannot be determined without some extraneous information.

The approach outlined in this chapter is essentially one of functional analysis, but applied to more than one set of readings. The extraneous information is then provided by making the equation fitted to one set of data go through the means of *mother set* of data.

The Wald–Bartlett Approach. Another form of functional analysis of a single set of data is to divide the data either into two sub-groups (Wald, 1940) or three sub-groups (Bartlett, 1949). Then an equation can be fitted to the sub-group means essentially along the lines of this chapter.

However, dividing a single set of data into two or three sub-groups is an arbitrary matter if no *external* criterion is being used (such as age, race, sex, etc. in the height/weight example). The resulting equation therefore does not carry any of the connotations of an empirical generalisation and the Wald-Bartlett methods do not seem to have caught on in practice.

Instrumental Variables. This is a method discussed mainly in the literature of econometrics. It arises when there is a third variable, like *age* in our height and weight example. For example, in Exercise 71 we derived the equations $h = 1.75A + 37$ between height and age, and $w = 5.25A + 22$ between weight and age. Eliminating the age-variable A from these two equations (by writing $(h - 37)/1.75 = (w - 22)/5.25$) gives the equation $w = 3h - 89$: virtually the same height/weight working-solution as in Section 7.6. The age-variable has therefore been "instrumental" in deriving this relationship.

This process is discussed more generally in Chapter 10, as part of the general use of theoretical arguments. The analytic approach described in this chapter has much in common with the broad idea of instrumental variables. But in the literature not much emphasis-is placed on the property of empirical generalisation, that the same height/weight equations holds despite the different values of the instrumental variable.

CHAPTER 8

Non-linear Relationships

Most empirical relationships require non-linear mathematical functions to describe them. The purpose of this chapter is to start to bridge the gap between fitting initial linear working-solutions and deriving more complex curvilinear models.

Our discussion will largely centre on the height/weight relationship. As more data are considered, we will see how the initial linear working-solution fitted in Chapter 7 develops into the logarithmic relationship referred to in Chapters 5 and 6.

8.1 Systematic Deviations

We finished our analysis of the heights and weights of privileged Ghanaian boys in the last chapter with several possible solutions and had some doubts about which one to use. Such doubts are common when working with initial solutions based on limited evidence.

The possibilities included the linear equation $w = 3.2h - 100$ and Curve C, shown again in Figure 8.1. Since in both cases the deviations are not large, one would normally choose the straight line because it is simpler to use. But we can see in Figure 8.1 that a linear equation like $w = 3.2h - 100$ could not hold much outside the range of variations covered so far: it says, for example, that children standing 30 inches or less have no weight.

Therefore some kind of curve will ultimately be needed to integrate the wide range of data for children of other ages. But at this stage we have no indication of which kind of curve to fit and have to look at more data.

8.2 More Data in the Same Range

Table 8.1 gives readings taken in 1947/8 for 5- to 13-year-old boys from Birmingham, England, of the middle Social Class "3" (Healy, 1952; Clements, 1953; Ehrenberg, 1968.) These cover about the same range of variation as the Ghanaian boys.

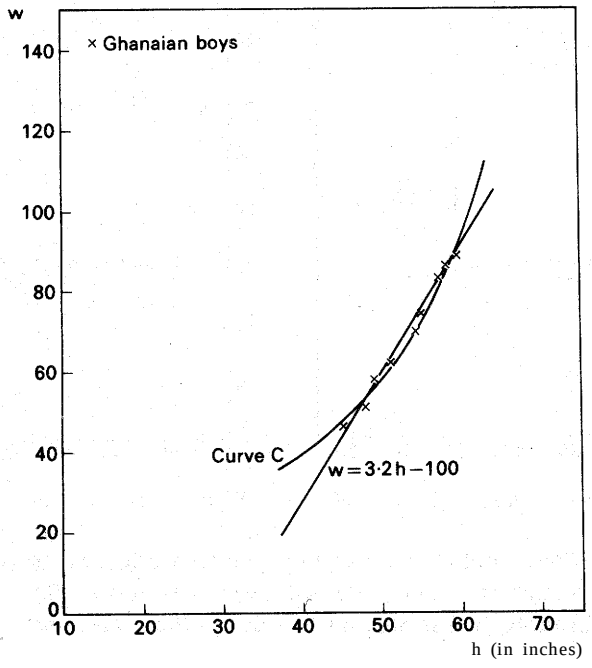

Figure 8.1 Two Working-Solutions for the Ghanaian Boys

The deviations of these data from the linear working-solution $w = 3.2h - 100$ are somewhat larger than for the privileged Ghanaian boys in the last chapter: on average 3.1 lbs compared with 1.3 lbs. They are also mostly negative and have some kind of pattern in that the two extreme values are either positive or small. This suggests that a curve would fit the Birmingham data better than any straight line, and supports the earlier suggestion in

TABLE 8.1 The Fit of the Previous Working-Solution w = 3.2h − 100 for Birmingham Boys of Social Class (3) Aged 5 to 13 Years

Birmingham Boys	Age 5	6	7	8	9	10	11	12	13	Av.
Av. height: h (ins.)	43	46	48	50	52	54	55	58	59	51.7
Av. weight: w (lbs.)	42	46	51	57	62	68	73	82	88	63.2
3.2h − 100	38	47	54	60	66	73	76	86	89	65.4
w − (3.2h − 100)	4	-1	-3	-3	-4	-5	-3	-4	-1	-2.2*

* Average size ignoring sign = 3.1 lbs

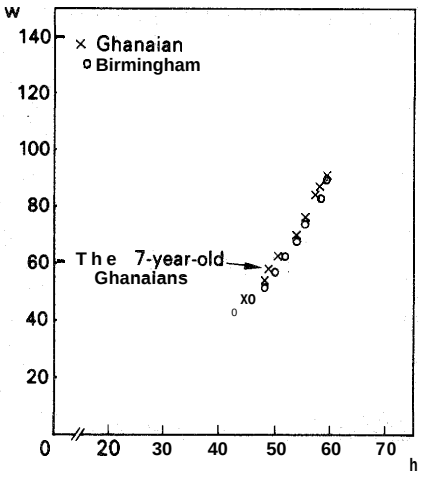

Figure 8.2A Consistent Curvature?

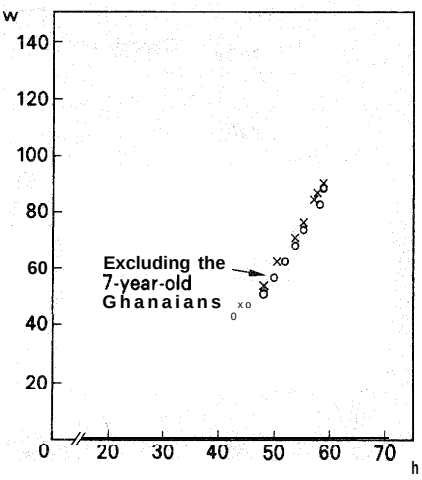

Figure 8.2B Excluding One Reading

the Ghanaian data that the relationship generally might be curved. The next question is whether the same curve will serve both sets of data.

Figure 8.2A plots the Birmingham results on the same graph as the results for the privileged Ghanaian boys. It shows that both sets of data are fairly similar. In Table 8.2 the two sets of deviations from the equation $w = 3.2h - 100$ are compared numerically. On average the Birmingham boys are about 2 lbs lighter, but this difference is small compared with the nearly 50 lbs total range of weights. The pattern that emerges still confirms the suggestion of curvature; not a simple $+-+$, but a positive deviation at 5 years changing to large negative ones at 8 to 10 years, and then back on average to *small* negative ones at the higher ages.

TABLE 3.2 The Deviations from the Linear Working-Solution $w = 3.2h - 100$ for the Birmingham and the Ghanaian Boys

The Deviations ($w - 3.2h + 100$)	Age 5	6	7	8	9	10	11	12	13	Av.
Birmingham boys	4	-1	-3	-3	-4	-5	-3	-4	-1	-2
Ghanaian priv. boys	2	-2	1	-1	-3	-1	1	0	-1	0
Average	3	-1	-1	-2	-3	-3	-1	-2	-1	-1

Sometimes a single reading can wrongly dominate an analysis. Looking at Figure 8.2A it is apparent that the 7-year-old Ghanaian boys are relatively heavy. They have a positive deviation of 1 lb in a context of negative deviations for boys aged 6 to 8 years. If we exclude this reading, the similar curvature of the two sets of data becomes more apparent, as in Figure 8.2B.

This might seem like "subjective messing around with the evidence", but one is only noting possibilities, e.g. that the same curve might describe both sets of data. The crucial next step is finding that in fact this possibility has also been supported by a wide variety of other data, as has been summarised in Table 5.1.

This need not have happened. We could have found something like the notional data in Figure 8.3. Here the deviations from the straight line are sometimes positive and sometimes negative, even at the same point on the line. They do not fall into a simple pattern. A straight line therefore gives the simplest summary that any *single* equation in the two variables could provide.

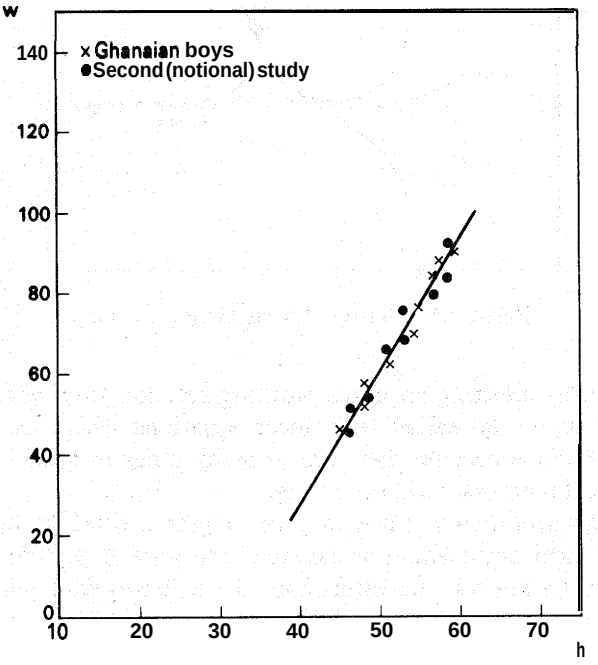

Figure 8.3 Inconsistent Deviations (Notional Data)

Another possibility is that different sets of data would not agree even about the broad nature of the relationship. This would lead to a complete failure to generalise, as was illustrated in Chapter 6.

8.3 **Choosing a Curve-Transforming One Variable**

Having seen that the height/weight data follow a curved relationship, we have to decide what kind of curve to fit. Rather as we saw with straight

lines, a variety of different mathematical functions can closely approximate any particular set of data. This is especially true when there is a limited range of variation and only a slight degree of curvature in the data. But the alternative curves may differ greatly outside the observed range of variation, as Figure 8.4 shows. Ultimately, this will help us choose which to use.

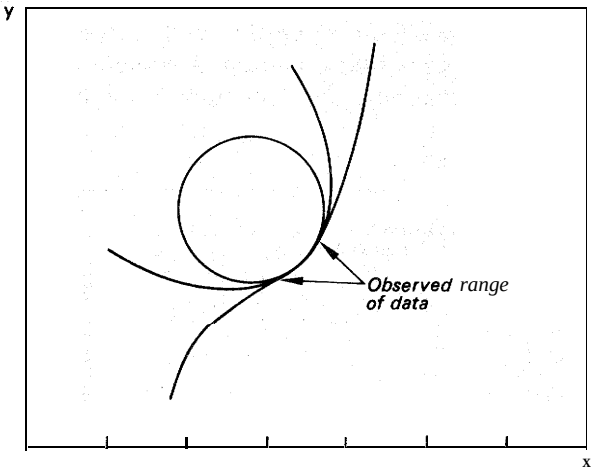

Figure 8.4 Different Curves Giving a Good Fit

We begin by selecting an initial working-solution from various possible curves, just as we did earlier with linear equations. This time the problem of choice is worse because there are so many different kinds of non-linear mathematical functions to choose from.

Sometimes a particular curve may be suggested either by an analogy or by a theoretical argument. For example, the growth pattern of a certain animal may be known, suggesting that the height/weight relationship for boys is of the same mathematical form. But in the absence of such outside knowledge, the usual procedure is to see whether the curved relationship can be dealt with by changing it into a linear equation. This is done by "transforming" the scale of measurement of one or both of the variables.

There are no clear rules for finding a suitable mathematical transformation. It is usually a matter of trial and error. The procedure depends on seeing the "shape" of the observed data and being aware of a variety of mathematical functions which might change that shape into a straight line. In practice, the analyst usually tries a small number of common mathematical functions, e.g. square roots, squares, reciprocals, exponentials, and logarithms. Other functions usually stem from a stronger theoretical basis (as illustrated in Chapter 10).

We saw in Figure 8.2 that the degree of curvature is small for the height and weight data. Transforming just one of the variables should therefore suffice. To help decide which one, we have to look for some extra evidence or suggestions. These usually exist, even if they are weak.

In our case, some evidence is provided by the separate relationships of age with height and weight. Between height and age the relationship is more or less linear, but between weight and age it is less so (see Exercise 71). This suggests that a suitable transformation of *weight* should straighten out the relationship of weight with height and also that of weight with age, a beneficial side-effect. There is nothing mandatory about trying this approach, but at this early stage of study the suggestion is enough to make one explore it.

Next we have to choose an actual transformation for weight. Given the shape of the data in Figure 8.2, we are seeking to shorten the intervals between large weights. This will have the required effect of "pushing down" the top-end of the height/weight curve and straightening it out.

Squaring the weights would have the wrong effect since large values would be spread out more than before. Equally spaced values of 1, 2, 3, 4 when squared become 1, 4, 9, 16, with increasing gaps of 3, 5, 7. In contrast, a *square-root* transformation would have the required kind of effect. Values of 1, 2, 3, 4 would become 1.00, 1.41, 1.73, 2.00, with *decreasing* intervals of .41, .32, .27. Such a transformation would work with the height/weight data but has nothing special to recommend it.

Using the *logarithm of* weight would have the same desired effect. It also has a slightly special appeal because it fits in with certain broader notions of biological growth functions. These often take logarithmic or exponential forms because growth tends to be multiplicative rather than additive. Things often grow *proportionally* to their size, e.g. by 10% a year rather than by a fixed amount per year. The argument is not a *strong* one when applied to weight. (Why not also to height, for instance?) But historically it largely led to a logarithm transformation being tried.

8.4 The Fit of the Logarithmic Relationship

Having decided to try a logarithmic transformation of weight, we next have to write down the log values of the average weights, using a table of logarithms. The 5-year-old Birmingham boys had an average weight of 42 lbs. The number given for 42 in a log table is .62 and since it is a number in the tens, a 1 has to be added and the log is 1.62. The results for the 5- to 13-year-old Birmingham boys are set out in Table 8.3. (We have to use three digits because the first ones are always the same and hence do not contain "significant" information.)

To fit a linear equation of the form $\log w = ah + b$, we use the same procedure as before. From the extreme readings for the 5- and 13-year olds

TABLE 8.3 The Logarithmic Values of the Average Weights of the Birmingham Boys

Birmingham Boys	Age 5	6	7	8	9	10	11	12	13	Av.
Av. weight:										
w (lbs.)	42	46	51	57	62	68	73	82	88	63.2
log w (log lbs.)	1.62	1.66	1.71	1.76	1.79	1.83	1.86	1.91	1.94	1.79

we get the slope-coefficient

$$a = \frac{1.94 - 1.62}{59 - 43} = \frac{.32}{16} = .020.$$

Inserting the overall averages of 1.79 and 51.7 into the equation, we get the intercept-coefficient

$$b = 1.79 - .020 \times 51.7 = .76.$$

Therefore the working-solution is

$$\log w = .020h + .76.$$

Table 8.4 shows the fit of this equation, which gives a mean deviation of .01 log lb units.

TABLE 8.4 The Fit of the Working-Solution log w = .020h + .76

Birmingham Boys	Age 5	6	7	8	9	10	11	12	13	Av.
Av. height: h	43	46	48	50	52	54	55	58	59	51.7
Av. weight: log w	1.62	1.66	1.71	1.76	1.79	1.83	1.86	1.91	1.94	1.79
.020h + .76	1.62	1.68	1.72	1.76	1.80	1.84	1.86	1.92	1.94	1.79
log w - .020h - .76	.00	-.02	-.01	.00	-.01	-.01	.00	-.01	.00	-.01*

* Average size ignoring sign = .01 log lbs

The logarithmic transformation has worked in the sense of yielding a simple linear relationship between log weight and height for the Birmingham data. Figure 8.5.A illustrates the relationship as a curve when plotting weight in lbs against height. Figure 8.5B shows it as a straight line when plotting log w against height.

The deviations, although mostly negative, show no trend. The logarithmic transformation has therefore succeeded in accounting for the curvature of the original data. The tendency for negative or zero deviations in Table 8.4 means that an equation with an intercept-coefficient of .75 would fit better.

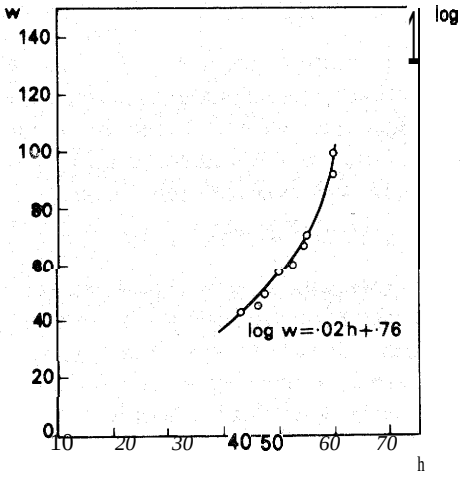

Figure 8.5A A Logarithmic Relationship

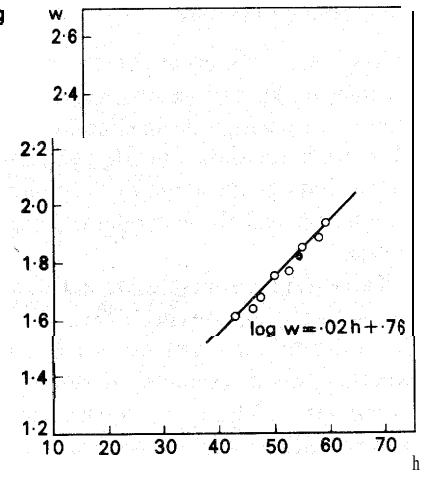

Figure 8.5B Log *w* against *h*

But the negative deviations are merely due to rounding-off. We could have eliminated the problem by working to three decimal places when calculating ***b.*** The average of the log w values is 1.787, and the average of the theoretical estimates is 1.793; these differ by $-.006$, which is $-.01$ to two places, but both numbers are 1.79 when rounded. This illustrates the rather niggling type of problem that can arise when working to only two significant figures but which is usually worth putting up with. (The real question is which value, .75 or .76, works better with *further* data.)

The next step then in the analysis is to see whether the relationship log w $= .02h + .76$ generalises to other data. We start with readings in about the same height and weight range as the Birmingham and Ghanaian boys. The results of such wider checks have already been summarised in Table 5.1. Table 8.5 gives an example for Canadian girls. The fit is good again, to within $\pm .01$.

TABLE 8.5 The Fit of the Working-Solution log w = .020h + .76 for Canadian Girls Aged 6 to 11 of Above -Average Socio-Economic Class

Ottawa Girls	Age 6	7	8	9	10	11	Av.
Av. height : h	47	48	51	52	54	56	51
Av. weight : log w	1.68	1.72	1.77	1.81	1.84	1.89	1.79
.020h + .76	1.70	1.72	1.78	1.80	1.84	1.88	1.79
log w - .020h - .76	-.02	.00	-.01	.01	.00	.01	.00*

*Average size ignoring sign = .01 log lb

8.5 Strong Curvature

The main doubt about the working-solution log w $= .02h + .76$ is whether we have chosen the correct function to transform the weights. To subject the logarithmic function to a stringent test, we have to see whether it also holds for data outside the range of variation covered so far. Since we know that *age* is the major determinant in the values of children's height and weight, we test the log relationship with data outside the 5- to 13-year age range.

Table 8.6 gives height/weight data for Ghanaian pre-school children aged 0 to 4 years (Kpedekpo, 1971). The logarithmic relationship fits well for the 2- to 4-year olds, but not for babies aged one year or less. It is not clear whether this discrepancy is general for babies, something specific for these Ghanaian children, or a measurement bias, because this is the only data for young children and babies analysed so far.

TABLE 8.4 Pre-School Children and the Working-Solutions
log w = .02h + .76 and w = 3.2h - 100

Pre-School Ghanaian Children	Age 0	1	2	3	4
Av. height : h	24	30	33	37	39
Av. weight : log w	1.16	1.32	1.42	1.48	1.54
.02h + .76	1.24	1.36	1.42	1.50	1.54
log w - .02h - .76	-.08	-.04	.00	-.02	.00
Av. weight : w	14	21	26	30	35
3.2h - 100	-25	-4	6	18	25
w - 3.2h + 100	39	25	20	12	10

Nevertheless we can see from Table 8.4 that the logarithmic relationship fits these data vastly better than the linear working-solution $w = 3.2h - 100$. (For the 5- to 13-year olds the logarithmic equation was only *fractionally* better, in the sense that it only had to account for the minor curvature in the data.)

To test the new relationship for older children, we look at data for some teenage children up to 17 years (Lovell, 1972). Table 8.7 shows that the logarithmic relationship holds well for the boys, but the deviations for the girls are all positive and about .04 log lb, much larger than average. Other data (Lovell, 1972; Kpedekpo, 1970) have confirmed the finding that older girls are relatively heavy for their height (therefore the logarithmic relationship itself cannot be blamed for these discrepancies).

TABLE 8.7 Older Boys and Girls and the Relationship log w = .02h + .76

Prosperous African Jamaicans	Age 14	15	16	17	Av.
BOYS					
Av. height : h	64	66	68	69	66.8
Av. weight: log w	2.04	2.08	2.12	2.13	2.09
.02h + .76	2.04	2.08	2.12	2.14	2.09
log w - .02h - .76	.00	.00	.00	-.01	.00
GIRLS					
Av. height : h	63	63	63	64	63.2
Av. weight: log w	2.06	2.07	2.07	2.07	2.07
.02h + .76	2.02	2.02	2.02	2.04	2.03
log w - .02h - .76	.04	.05	.05	.03	.04

We now no longer have any doubts about whether the height/weight data need to be described by a non-linear equation. Had we started our study with a wider age-range, we would at once have been faced with a strongly curved empirical relationship, like the one shown in Figure 8.6A. Then we would not even have tried a straightforward linear solution. We would probably have transformed one of the variables immediately in order to reach the type of linear situation shown in Figure 8.6B, which is easier to analyse. That way

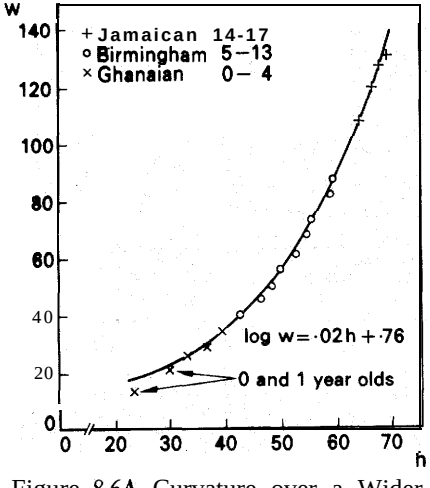

Figure 8.6A Curvature over a Wider Range

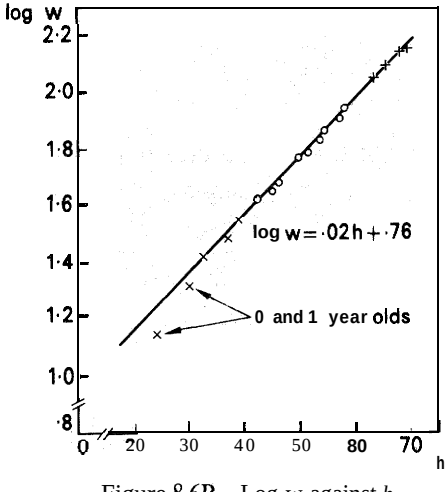

Figure 8.6B Log w against *h*

we could have determined more quickly whether the right kind of model had been chosen.

8.6 A Cube-root Law

The fact that the logarithmic equation fits well for boys aged 2 and over does not eliminate other possible mathematical functions. For example, a cube-root transformation can also be considered.

There is a dimensional argument in favour of using a cube-root equation. We know that the weights of different bodies vary with their volumes, and that the volume of a right-angled body varies as the product of three linear dimensions: height, width, and depth. It follows that the cube-root of volume, and hence the cube-root of weight, should be proportional to height. We can therefore try to fit an equation of the form

$$\sqrt[3]{w} = ah.$$

However, a check on the data shows that the variation is not directly proportional and that the more general equation $\sqrt[3]{w} = ah + b$ would work better. This does not mean that the theoretical argument is already being discarded? only that it is being shaped to fit the facts. This is legitimate because the deduction that height should vary directly with the cube-root of weight was not cast-iron anyway; it assumed that the shape of children was rectangular, and that this and their density would remain exactly the same as they grew. Obviously some adjustments are permissible.

Table 8.8 gives the height and weight readings for boys aged 0 to 17 (covered in Tables 8.4, 8.6 and 8.7). Fitting a straight line for $\sqrt[3]{w}$ and h in the usual manner gives

$$\sqrt[3]{w} = .060h + 1.0.$$

TABLE 8.8 The **Cube-Root** Relationship $\sqrt[3]{w} = .060h + 1.0$

(Ghanaian children aged 0 to 4, Birmingham boys aged 6 to 12, and Jamaican boys aged 14 to 17)

	Age													
	0	1	2	3	4	6	8	10	12	14	15	16	17	Av.
Av. height: h	24	30	33	37	39	46	50	54	58	64	66	68	69	48.9
Av. weight: w	14	21	26	30	35	46	57	68	82	110	122	131	134	67.4
$\sqrt[3]{w}$	2.4	2.8	3.0	3.1	3.3	3.6	3.8	4.1	4.3	4.8	5.0	5.1	5.1	3.9
.060h + 1.0	2.4	2.8	3.0	3.2	3.3	3.8	4.0	4.2	4.5	4.8	5.0	5.1	5.1	3.9
w - .060h - 1.0	.0	.0	.0	-.1	.0	-.2	-.2	-.1	-.2	.0	.0	.0	.0	.0*

* Average size ignoring sign = .1

This equation fits the data to within about 0.1 "cube-root lb" units. All the deviations are small. There are negative deviations in the middle of the range; but we cannot tell whether this generalises until we investigate additional data.

What matters is that the cube-root equation works almost as well as the logarithmic one for describing the height/weight relationship of children aged 2 years and over, and that it ***also*** gives a close fit for babies aged one year or less, as is also shown in Figure 8.7. This suggests that perhaps babies are not different from older children, but that it was a failure of the earlier theory, the logarithmic transformation, that caused the marked overestimation in Table 8.6. Further work and more data are still needed to determine this.

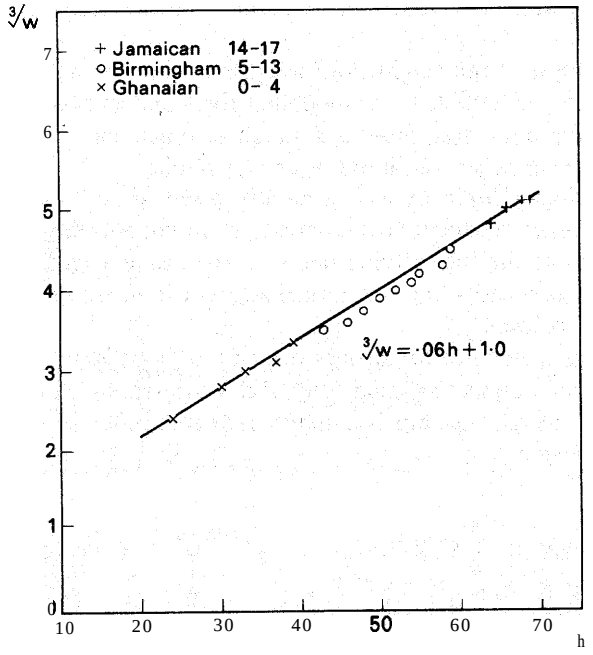

Figure 8.7 A Cube-Root Relationship

More generally, the success of the cube-root transformation so far means that our initial "theoretical deduction" has connected these results to our prior knowledge about the dimensionality of height and weight. This integrative function of theory is often far stronger than here. More examples will be discussed in Chapter 10.

8.7 Summary

When modelling a non-linear relationship, the main problem is choosing an appropriate mathematical function. Because this can be a complex operation, it is usually worth establishing first that the curved pattern generalises to other sets of data.

Usually a number of different curves or mathematical functions can approximate any particular set of data, especially when the range of variation is limited. Previous work, a theoretical argument, or an analogy can help suggest which functions to try.

Sometimes little previous information is available. Then a simple transformation of one of the variables is usually tried to lead once more to a linear type of working-solution. Unless there is some well-based theory this becomes a matter of trial and error. One has to see the shape of the observed data and be aware of a variety of mathematical functions that could meet that shape.

In our example of the heights and weights of children aged 5 to 13 years, the equation $\log w = .02h + .76$ modelled the slight curvature in the data. It also held for data that showed a much stronger curvature outside the initial range of variation, but it did not fit for babies.

While the logarithmic fit was generally good, it did not preclude the possibility of other mathematical functions. A cube-root transformation held almost as well as the logarithmic one and even gave a good fit for babies. This function also had some theoretical support in terms of the dimensions of weight and volume.

At this stage the choice between two such functions is not very important because they tell almost the same story about the available data. The nature of the functional relationship is usually clarified further as more advanced theory develops.

CHAPTER 8 EXERCISES

Exercise 8A. A Curved Relationship

Table 8.9 gives measurements of the apparent brightness of a light-source at various distances from it. Fit an equation to summarise the data.

TABLE 8.9 *The* Brightness of a Light-Source at Various Distances

Distance D (ft.)	1	2	3	4	5	6	7
BrightnessB(lumens)	1900	530	210	120	80	62	21

Discussion.

It is obvious from the table (or from a rough working graph) that brightness ***B*** decreases as distance ***D*** increases. It is also clear that ***B***

decreases at a much faster rate at short distances. For example, brightness at 4 ft, half the distance covered, is far less than half of the range of B-values observed.

The relationship is very highly curved, so that a transformation of at least one variable is needed. Since distance in feet is a well-known and highly controllable measure, it might seem best to leave this alone and transform brightness. Logarithms, as in Table 8.9a, would largely straighten out the relationship, but not entirely. For example, log B drops by 1.20 from 1 to 4 ft, but by only .76 from 4 to 7 ft.

TABLE 8.9a Logarithm6 of Brightness

Distance as D	1	2	3	4	5	6	7
Brightness as log B	3.28	2.72	2.32	2.08	1.90	1.79	1.32

To hit on a better transformation we can use the physical law that brightness and certain other measures, such as the force of gravity, vary inversely with the square of distance, i.e. as $1/D^2$ (so that we transform D after all).

Table 8.9b sets out the original brightness readings with the reciprocal of the squared distances D, multiplied for numerical convenience by 1,000.

TABLE 8.9b B and $1000/D^2$

Distance as D	1	2	3	4	5	6	7	Av.
Brightness as B	1900	530	210	120	80	62	21	418
Distances as $1000/D^2$	1000	250	110	62	40	28	20	216

It is obvious that B is generally about twice as large as $1{,}000/D^2$, so that the relationship is something like

$$B = \frac{2{,}000}{D^2}.$$

The main discrepancy is at 7 feet, where B is virtually equal to $1{,}000/D^2$, rather than $2{,}000/D^2$. This requires further study. Perhaps the measuring procedure used is relatively erratic at low brightness levels.

Exercise 8B. AMeasurementBias?

Table 8.10 shows that the relationship $B = 2{,}000/D^2$ does not fit data for the same light source at greater distances. Analyse the data further.

TABLE *8.10* Brightness at Greater Distances

Distance D (ft)	10	15	20	25	30	35	Av.
Brightness B (lumens)	15	9	7	7	6	6	8.3
$2000/D^2$	20	9	5	3.2	2.2	1.6	6.8

Discussion.

In general, the apparent brightness of a light source *does* vary inversely with the square of the distance. Since this formulation gives roughly the right pattern in Table 8.10, we shall keep to the $1/D^2$ function at this stage, but first fit a more general equation of the form

$$B = a\left(\frac{1{,}000}{D^2}\right) + b.$$

Determining a linear working-solution for B and $(1{,}000/D^2)$ in the usual manner gives the equation

$$B = \frac{1{,}000}{D^2} + 4.9.$$

(Note that Table 8.10 gives $2{,}000/D^2$, not $1{,}000/D^2$.) The new equation fits the brightness values with a mean deviation of ± 0.3 lumens.

A possible interpretation of this result is that the device for measuring brightness had a bias of about 4.9 or 5 lumens. The relationship might therefore read

$$(B - 5) = \frac{1{,}000}{D^2},$$

where $(B - 5)$ is the *real* level of brightness.

There could of course be other causes. For example, some light could be reflected from surrounding material, or other sources of light might be registered at lower brightness levels. Or it could be the *distance* measure that is biased, e.g. by being taken from the edge of the light source instead of its focus. In the latter case, the corrected form of the inverse-square law would read $B = \hat{a}/(D - \hat{b})^2$, where $\hat{a}$ and $\hat{b}$ are two coefficients to be determined.

Establishing the relationship in this form would be technically complicated because it cannot be expressed linearly by transforming ***B*** and/or *D*. But written the other way round, in the form $D = \hat{b} + \sqrt{\hat{a}}/\sqrt{B}$, the relationship is linear in *D* and l/,/B.

Determining an appropriate model for these phenomena and, in particular, reconciling these data and those from Exercise 8A, will clearly take more extensive data and a reasonable degree of theoretical understanding.

ExerciseSC. AChangeofUnits

The relationship between children's height ***h*** and weight w is log w = $.02h + .76$ in inches and pounds. What is it in metres and kilograms? (1 inch = 2.54 centimetres; 1 pound = 454 grams.)

Discussion.

If H is height in metres, then $100H/2.54$ or $39H$ equals h in inches. (H must be numerically smaller than h because there are fewer metres than inches in a given length.) Similarly, if W is weight in kilograms, then $1{,}000W/454 = 2.20W$ is equal to w measured in lbs.

It follows that log w = log $2.20W$. This in turn equals log 2.20 + log W because of the multiplicative property of logarithms (log XY = log X +

log Y). From log tables, log 2.20 = .342 so that

$$\log w = .342 + \log W.$$

Substituting for log w and h in the equation log w = .02h + .76 it follows that

$$.342 + \log W = .02 \times 39H + .76.$$

Hence

$$\log W = .78H + .42.$$

This is virtually the equation log $W = .8H + .4$ that was reported by Lovell (1972) and others, in metric units.

Similarly, the reader may check that changing the cube-root formulation $\sqrt[3]{w} = .060h + 1.0$ to metric units gives

$$\sqrt[3]{W} = 1.8H + .77.$$

Exercise 8D. Deviations from the Log Relationship

In Section 8.4 the mean deviation from the relationship log w = .020h + .76 was ± .01 log lbs. What is the equivalent deviation in ordinary pound units?

Discussion.

There is no single answer. Constant limits of ± .01 log lbs correspond to differing values in pound units, depending on the absolute value of w. Thus the deviations in lbs are larger for large values of w than for small ones.

To illustrate, consider three values of log w in about the range covered by the data for the 5- to 13-year olds, namely 1.60, 1.75, and 1.90. The values of log w ± .01 are

log(w) + .01:	1.61	1.76	1.91
log (w) :	1.60	1.75	1.90
log(w) − .01:	1.59	1.74	1.89.

Looking up the antilogarithms of 1.61, etc., the equivalent values of weight w in pound units are

w equivalent to log(w) + .01:	40.7	57.5	81.3
w equivalent to log (w) :	39.8	56.2	79.4
w equivalent to log(w) − .01:	38.9	54.9	77.6

This shows that deviations of ± .01 in log lb units are equivalent to about ± 1 lb at the 40 lb level, ± 1.3 lbs at the 56 lb level, and about ± 1.8 lbs at the 80 lb level.

When an empirical relationship between two variables x and y is non-linear and curved upwards as here, deviations from it in the original units tend to be larger for larger values of x or y. When the relationship is "linearised" by some suitable transformation, the deviations from the resultant straight line, i.e. deviations in the transformed units, tend then to be more or less homoscedastic (i.e. have equal scatter) all along the line.

This shows up here particularly in terms of the larger scatter of *individual* children about the line. Thus for the linear equation w = 3.2h − 100, the mean deviation of 8 lbs quoted in Section 7.8 was a broad average, there

being a marked upward trend from 5 to 13 years. But in log units the scatter was virtually homoscedastic at .04 log lbs, as quoted in Section 6.4.

A related technicality is that if one first transforms the *individual* weights to logs their average will be slightly smaller than the log of the average weights in lbs that we have used. (This is related to the well-known difference between arithmetic and geometric means.) This can matter when working at high levels of precision.

Exercise 8E. TheFitinLogarithmicandAdditiveUnits

The fit of the age-group averages to log $w = .02h + .76$ was on average within .Ol log lb units and that for the linear equation $w = 3.2h - 100$ in Section 7.7 was about 1.2 lb. Is the fit of the logarithmic equation closer?

Discussion.

Exercise 8D showed that deviations of $\pm .01$ in log lb units are equivalent to deviations of $\pm$ 1.3 lbs at the mid-range of w-values, but *smaller* than 1.3 for lower values of w and *larger* for higher values. No single "neat" comparison is therefore possible. But as a broad average, the two deviations must be roughly the same, because both equations gave a fair fit to the same data (i.e. for the 5- to 13-year olds in Table 8.3).

As a *proportion* of the weight values the two average deviations differ. From the previous Exercise we know that $\pm$ 1.3 of 60 lbs (the mid-range w-values) is about 2%, whereas $\pm .01$ of 1.8 log lbs is about 6%. But such a comparison is not very meaningful, as it depends on the rather arbitrary zero values of the two scales. A more appropriate comparison would be the size of the deviations compared to the *range of variation* on each scale. Here we have

A range of .34 in log lbs, so that $\pm .01$ log lbs is about $\pm$ 3%,

A range of 40 in lbs, so that $\pm$ 1.3 lbs is also about $\pm$ 3%.

However, the crucial difference between the two equations is not the *size* of the deviations, but the fact that the deviations for the relationship $w = 3.2h - 100$ are systematic, especially outside the range of the 5- to 13-year olds (see Table 8.6).

ExerciseSF. Children's Weight and Age

In Exercise 71 we noted that the relationship between weight and age for the 5- to 13-year-old Ghanaian boys appeared to be slightly non-linear. Examine this further.

Discussion.

One way of dealing with the weight-age curvature is a logarithmic transformation of weight. Table 8.11 shows that this also works well for 5- to 13-year-old Birmingham boys, with at most a slight pattern in the small residuals.

A stronger test for the equation would be a check outside the above range, like the data on black Jamaican teenagers in Table 8.7. The girls tend to be taller at a given age than the relationship indicates. But the difference is consistent (about .05 log lb units) at various ages, so the

TABLE 8.11 The Fit of the Working-Solution log w = .04A + 1.43

Birmingham Boys	Age in Years (A)									Av.
	5	6	7	8	9	10	11	12	13	9.0
Av. weight: log w	1.62	1.66	1.71	1.79	1.79	1.83	1.86	1.91	1.94	1.79
.04A+1.43	1.63	1.67	1.71	1.75	1.79	1.83	1.87	1.91	1.95	1.79
log w - .04A - 1.43	-.01	-.01	.00	.01	.00	.00	-.01	.00	-.01	.00

logarithmic transformation has in fact succeeded in "straightening out" the relationship. More extensive study is needed to clarify how average heights differ at a given age. This will be touched on again in Chapter 9.

Exercise 8G. AFailuretoFit

Table 8.12 compares the average weights of British boys from a survey in 1880 (Clements, 1953) with the equation log w = .04A + 1.43 fitted to the Birmingham boys measured in 1947. Comment on the data.

TABLE 8.12 The Deviations from the Birmingham 1947 Equation log w = .04A + 1.43 for Boys in 1880

Boys in 1880	Age in Years (A)									Av.
	5	6	7	8	9	10	11	12	13	9.0
Av. weight: log w	1.57	1.61	1.67	1.71	1.79	1.78	1.81	1.89	1.88	1.73
.04A+1.43	1.63	1.67	1.71	1.75	1.79	1.83	1.87	1.91	1.95	1.79
log w - .04A - 1.43	-.06	-.06	-.04	-.04	-.05	-.05	-.06	-.07	-.07	-.06

Discussion.

The Birmingham equation does not fit, but the deviations are consistent at about −.06 log lbs. The boys in 1880 therefore weighed markedly less at any given age than those in 1947. The equation

$$\log w = .04A + 1.37$$

would be a good working-solution for the earlier data.

Exercise 8H. Changing the Mathematical Function

In Section 8.6 we changed from a logarithmic to a cube-root function of weight. This is typical of the change in mathematical functions that can occurs as a subject develops. When doing this, is it necessary to refer again to the original data?

Discussion.

No. Unless more precision is required than the average deviations of ± .Ol log lbs, the theoretical relationship log $w = .02h + .76$ summarises the original data sufficiently and no direct reference back is needed. (This is like the change in linear relationships discussed in Exercise 7F.)

To examine the fit of a cube-root relationship, we can reconstruct the data from the logarithmic working-solution, as shown in Table 8.13.

TABLE 8.13 Deriving and Testing a Cube-root Relationship from the Logarithmic Relationship

Height h (in.)	Selected Values 20	40	60	80	Av. 50
.02h + .76 = log w	1.16	1.56	1.96	2.36	1.76
w (in lbs)	14	36	91	229	92
$\sqrt[3]{w}$	2.4	3.3	4.5	6.1	4.1
.060h + 1.0	2.2	3.4	4.6	5.8	4.0
$\sqrt[3]{w}$ - .06h - 1.0	.2	-.1	-.1	.3	.1

For a systematic selection of h-values covering the relevant range, we work out $.02h + .76$. This gives the theoretical values of log w and we look up antilogarithms to find the values of w. Then we calculate $\sqrt[3]{w}$ and fit a linear equation to h and $\sqrt[3]{w}$. This gives $\sqrt[3]{w} = .060h + 1.0$. The final step is to check the fit of the new equation. (More elaborate calculations can also be made using the $\pm .01$ average limits of approximation of the logarithmic relationship.)

CHAPTER 9

Many Variables

When analysing data with three or more variables, one usually starts by looking at two variables at a time. Many of the procedures therefore remain the same as in preceding chapters.

The main difference is that by interrelating the different paired relationships in the data, one can gain a fuller interpretation and more advanced theoretical structures.

9.1 An Initial Analysis: Apple Trees

To illustrate multivariate data at the earliest stages of analysis, we consider an example provided by Dr. S.C. Pearce (for a meeting of the Biometric Society in London in 1969). The data consist of four measures of the size of apple trees:

g = girth of trunk after 4 years,
G = girth of trunk after 15 years,
e = extension growth (length of branches, etc.) after 4 years,
W = weight above ground after 15 years.

Eight trees from each of 13 different rootstocks were planted in 1918/19. (The trees were Worcester Pearmains grafted on to rootstocks raised asexually.) It is regarded as a classical experiment in elucidating the effects of different rootstocks.

To see the pattern in the results we shall concentrate on the 13 rootstock averages for the four variables, instead of the readings for the individual trees. These averages have been arranged in Table 9.1 according to the size of G (before rounding). We chose G because girth is easier to measure than the other variables (and hence is more likely to recur in other data) and because girth *at 15 years* is probably a more stable measure than girth at 4 years. (This initial arrangement of the data is essentially arbitrary. One must check later whether any conclusions have been unduly influenced.) Even without rearranging rows and columns we can see from the table that the other three variables are quite highly correlated with G and that the

TABLE **9.1** The Average Values for the **13 Rootstocks,** in Increasing Order of Girth G at 15 **Years**

	Rootstocks													
	IX	VII	VI	I	x	IV	XIII	v	II	III	XV	XVI	XII	Av.
G, in cm.	24	33	36	37	37	39	39	43	45	45	45	45	48	40
e, in m.	17	29	22	30	30	29	29	23	31	28	36	40	47	30
g, in cm.	9	10	10	11	11	11	11	11	12	11	12	12	14	11
W, in **100 lb.**	4	7	7	9	8	10	11	12	13	14	13	17	19	11

relationships are roughly linear. (For example, the values for Rootstock XIII are very roughly mid-range for all four variables,) By using the normal procedure to fit working-solutions between G and each of the other variables. we obtain

$$G = e + 10,$$

$$G = 4.8g - 14,$$

$$G = 1.6W + 22.$$

The deviations from each relationship are shown in Table 9.2. They are mostly irregular and average at 3 cm, which is fairly small compared with G's range of 24 cm.

TABLE 9.2 The Deviations between the **Observed** and **Estimated** Values of G

	Rootstocks													
	IX	VII	VI	I	X	IV	XIII	V	II	III	XV	XVI	XII	Av.
G - e - 10	-3	-6	4	-3	-3	-2	0	10	4	7	-1	-5	-9	0
G - 4.8g + 14	-5	-1	2	-2	-2	-2	0	4	1	6	1	-1	-5	0
G - 1.6W - 22	-4	0	3	1	2	-1	-1	2	2	1	2	-4	-4	0
Av. size ign. sign	4	2	3	2	2	2	0	5	2	5	1	3	6	3

The all-negative or zero deviations for Rootstocks IX, VII, XVI and XII (at the two extremes of the table) suggest that the three relationships with G might be curved. There are also exceptional deviations in the relationship with *e* for Rootstocks V and III, and in the relationship with g for Rootstock III. Further data is needed to pursue these various deviations.

But the three relationships provide an adequate first summary of the given data. They also imply the relationships between the other pairs of variables. For example, from the two equations $G = e + 10$ and $G = 4.8g - 14$, we can deduce

$$e + 10 = 4.8g - 14,$$

so that

$$e = 4.8g - 24.$$

Relationships between g and W and between e and W can be derived similarly.

This is one of the crucial points about the analysis of many variables. Increasing the number of variables need not greatly complicate the final results. With four variables there are six equations between pairs of variables, but one only needs three equations to summarise the data. With ten variables there are 45 equations between pairs of variables, but one only needs to describe nine of them explicitly, and so on.

A second point is that having more variables leads to stronger interpretation. There is more information on which to base one's judgement. We can see a possible general factor of *size* due to specific rootstocks; some root-, stocks produced trees which tended to be bigger than others for all the variables measured.

A third point is that having even more data usually simplifies the results. We could reach firmer conclusions on any general patterns and on the exceptions if we had data on these trees at other ages and for additional variables (e.g. root growth, weight of apples cropped, etc.), and data for other trees. Since the results have been simple so far, it would not be difficult to integrate such additional data or to start asking ourselves what sort of additional variables we ought to try to analyse and what sort of results we would expect to find.

9.2 A More Structured Example: Consumer Attitudes

We now consider market research data on consumer attitudes as an example where a great deal of data are available (since many similar studies have already been made) and where the results all take the same general structure. It is also an example where we can start asking a few questions before going far in the analysis.

Market research surveys of consumer attitudes to frequently bought branded goods usually have a large number of variables. Potential consumers are asked whether they think each brand is modern, popular, sweet enough, convenient to use, good value for money, etc.: whatever seems relevant. Table 9.3 gives an extract of some typical attitudinal data for four brands of a certain product. The readings are the percentages of consumers who gave positive responses to the questions. For example, 36% of all consumers said Brand A has the "Right Taste".

The data look complex, but it is still possible to discern some order, such as the fact that Brand D scores very low for all the variables. We could now proceed as before and fit equations to pairs of variables. For example,

TABLE 9.3 The Percentage of Consumers Giving Certain Attitudinal Responsea to Brands A, B, C, and D

	% of All Consumers Saying:					
	"Right Taste"	"Convenient"	"Good Value"	"Popular"	"Modern"	etc.
Brand A	36	11	30	49	.	.
Brand B	11	4	40	9	.	.
Brand C	21	50	18	31	.	.
Brand D	6	3	4	6	.	.
etc.	.	.	.	.	.	.

between "Popular" and "Right Taste" we could fit the linear equation

$$P = 1.4RT - 2,$$

which gives a mean deviation of about 4 percentage points. But the analysis for the other variables is not as straightforward.

In any case, is this really the best way of summarising and interpreting such data? The readings refer to consumer attitudinal responses to branded goods, for which most of us have some background knowledge, so we can start asking questions. For example, *why* are the results for Brand D all so low?

One possible explanation is that Brand D is not very good, and that few people like it and use it. Another possibility is that Brand D *is* good, but either not well-known or not widely available (because of low advertising levels or limited retail distribution). This would *also* show up in few people actually using the brand.

We have two different interpretations, but both suggest attitudinal responses might be related to the number of consumers who use each brand. (We have already seen a similar connection in Chapter 4 for consumers' intentions-to-buy.) A simple measure of brand usage is available ("Have you used the brand in the last week?") and is introduced into the analysis in

TABLE 9.4 The Observed Attitudinal Responses and Usership Levels

	% of Consumers Using: U	% of All Consumers Saying: "Right Taste"	"Convenient"	"Good Value"	"Popular"
Brand A	50	36	11	30	49
Brand C	30	21	(50)	18	31
Brand B	10	11	4	(40)	9
Brand D	5	6	3	4	6
Average	24	18	6*	17*	24

* Averages excluding the Brand B and C exceptions

Table 9.4. Here the data are set out in decreasing order of each brand's market-share.

A pattern now becomes apparent. The attitudinal values decrease in parallel with the usership level. There are only two exceptions-for Brand C on "Convenient" and Brand B on "Good Value", (Reversing the columns and rows, as discussed in Chapter 1, would make the pattern even clearer.)

We cannot be sure of this pattern with just sixteen numbers, two of them exceptions. Again this is a case where more data would help rather than hinder. In fact, a wide range of other results exists (e.g. Collins, 1973; Bird and Ehrenberg, 1970; Chakrapani and Ehrenberg, 1974) and shows the same connection, that the level of such attitudinal responses to a brand is generally proportional to U, its user-level, but with some occasional highly dramatic exceptions.

Now we can start fitting working-solutions to the data in Table 9.4. But instead of trying to equate attitudinal variables to each other, as we would have done before, we shall look for relationships between each attitude and usership level, i.e. equations of the form

$$\text{Attitude} = aU + b.$$

The previous evidence suggests that the relationships have zero intercept-coefficients. Therefore, instead of using the highest and lowest readings to calculate the slope-coefficient a, it is simpler to use the ratio of the average for each attitudinal variable to the average market-share, across all four brands. Thus for "Right Taste" we have $a = 18/24 = .75$ or .8, rounded. (The two large exceptions in Table 9.4 are excluded from both averages. For example, for "Convenient" we calculate the slope as $a = 6/22 = .3$.) This gives the equations:

$$\text{"Right Taste"} = .8U$$
$$\text{"Convenient"} = .3U$$
$$\text{"Good Value"} = .6U$$
$$\text{"Popular"} = 1.0U$$

Table 9.5 shows the deviations between the observed and theoretical readings aU. The fit is close, with a mean deviation of about 2 percentage points.

We therefore have a much simpler model than in the case of the apple trees. The same form of relationship with usage holds for each variable. And as can be seen from the other published results, this form holds across a wide range of product-fields and different attitudinal measures.

We could still have analysed the data directly, relating the different attitude variables to each other. The above relationships in fact imply those between the attitudinal variables themselves. But the usage variable has

TABLE 9.5 Deviations between the **Observed** Attitude **Levels** and the Predicted **Values given by aU**

	"Right Taste" Obs. -.8U	"Convenient" Obs. -.3U	"Good Value" Obs. -.6U	"Popular" Obs. -1.0U
Brand A	-4	-4	0	-1
Brand C	-3	(41)*	0	1
Brand B	3	1	(34)*	-1
Brand D	2	1	1	1
Average	0	0	0	0

* Excluded from the calculations

introduced much more structure. It also links the attitudinal data to the surveys' real concern, i.e. consumers' actual usage and buying behaviour.

9.3 A **Simple Model: Buyer Behaviour**

Multivariate data can often appear extremely complicated, but may still be reduced to a simple, well-developed model. For example, individual consumers purchases of frequently bought branded goods are very irregular, as the data shown in Table 9.6 illustrate.

TABLE 9.6 Consumers' Purchasing of Different Brands of Breakfast **Cereals** in 12 Successive Weeks

(C: Corn Flakes, W: Weetabix, S: Shredded Wheat, P: Puffed Wheat)

	Purchases in Week											
	1	2	3	4	5	6	7	8	9	10	11	12
Consumer I	-	-	C	W	C	-	C	-	S	CS	-	C
Consumer II	P	-	-	-	-	-	-	-	-	P	-	-
Consumer III	-	W	W	-	W	-	W	-	W	W	-	-
Consumer IV	-	-	-	-	-	-	-	-	-	-	-	-
Consumer V	C	C	-	C	W	C	S	S	C	C	C	W
etc.	.	.	.	.	.	.	.	.	.	.	.	.

One approach to such data is to look at successive purchases, e.g. C, W, C, C, S, etc. for Consumer I. But consumers buy at different frequencies, so their purchases quickly get out of step with each other. Consumer II's second purchase follows Consumer I's *fifth.* This creates difficulties, especially when relating the data to outside events, such as seasons of the year, price changes, a new brand launch, advertising, etc.

An alternative approach is to consider specific time-periods such as successive 4-weekly ones, and to (examine separately

(i) the number of consumers 'who buy each brand at least once in each period, and

(ii) how often on average they buy it in the period.

Experience has shown that this approach leads to simple and generalisable results. This is the crucial test in choosing an analytic approach.

To explore the relationship between the different brands, suppose that we look at the number of consumers who buy any *pair* of brands in an analysis period. Table 9.7 gives an example of such data for five breakfast cereals in a 24-week period (from Charlton *et al.*, 1972). It shows that 35% of all consumers bought *both* Corn Flakes and Weetabix at least once in the period. (Each "duplication" percentage occurs twice in the table because the percentage of consumers who bought both Weetabix and Corn Flakes is the same as the percentage who bought Corn Flakes and Weetabix.) The diagonal shows each brand's "penetration", i.e. the percentage of all consumers who bought each brand at least once during the 24 weeks.

TABLE 9.7 **Brand-Duplication** of Purchase **in** 24 **Weeks**

(% of consumers buying any pair of brands, each at least once)

24 Weeks	and also buying				
	corn Flakes	Weet-abix	Shred. Wheat	Sugar Puffs	Puffed Wheat
Buying					
Corn Flakes	(61)	35	25	19	11
Weetabix	35	(49)	22	17	10
Shredded Wheat	2.5	22	(33)	11	9
Sugar Puffs	19	17	11	(26)	8
Puffed Wheat	11	10	9	8	(16)

The five brands have been arranged in decreasing order of their penetrations. This brings out a regular pattern : the duplication figures in each column (omitting the penetrations in the diagonal) decrease in the same order. Therefore the percentage of consumers who bought two brands varies with the penetration of each brand. More people bought Weetabix than bought Puffed Wheat (49 % versus 16 %), and more *Corn Flakes* buyers also bought Weetabix than Puffed Wheat (35 % versus 11%).

Now we have to establish the quantitative details to support this theory. One possibility is that the two penetration levels act independently of each other. If 49 % of the population buy Weetabix, then perhaps about 49 % of the *Corn Flakes* buyers also buy Weetabix. This gives 49% out of 61 %, or about 30 % who should buy both. It is close to the observed value of 35%, but a little below it.

Checking our model against the observed data in Table 9.7 shows that the Corn Flakes/Weetabix result generalises. The theoretical estimates are close to the observed values, but consistently lower. The deviations tend to be larger for the more popular brands, and this suggests adjusting the initial working-solution with a constant multiplier. Thus if b_X and b_Y represent the penetrations of Brands X and Y respectively, and b_{XY} is the percentage of consumers buying *both X* and Y, the adjusted model reads

$$b_{XY} = Db_Xb_Y/100,$$

where D is the average value of (b_{XY}/b_Xb_Y) across all pairs of brands. For the data in Table 9.7, D is about 1.3. (The divisor of 100 in the equation is needed because we are working in percentages, whereas in *proportions* the equation would read $b_{XY} = 1.3b_Xb_Y$.) Table 9.8 shows the fit of this model? the mean deviation being only 1.5 percentage points.

TABLE 9.8 The Fit of the Model $b_{XY} = 1.3b_Xb_Y$

(Observed and Theoretical Values 0 and T)

24 Weeks	Corn Flakes		Weet-abix		Shred. Wheat		Sugar Puffs		Puffed Wheat	
	O	T	O	T	O	T	O	T	O	T
Corn Flakes	-	-	35	39	25	26	19	21	11	13
Weetabix	35	39	-	-	22	21	17	17	10	10
Shredded Wheat	25	26	22	21	-	-	11	11	9	7
Sugar Puffs	19	21	17	17	11	11	-	-	8	5
Puffed Wheat	11	13	10	10	9	7	8	5	-	-
Average	23	25	21	22	17	16	14	15	10	9

Other mathematical formulations could also give a reasonable fit to this one set of readings. The importance of the expression $b_{XY} = Db_Xb_Y$ is that the same model holds for a wide variety of other time-periods and product-fields (e.g. Ehrenberg and Goodhardt, 1968, 1969; Ehrenberg, 1972). The same type of model also holds for other choice phenomena, such as industrial contracts (Ehrenberg, 1974b) and viewers' selection of television programmes (e.g. Goodhardt, 1966; Ehrenberg and Twyman, 1967; Goodhardt *et al.*, 1975). The theoretical model's slight overstatement of the observed results involving Corn Flakes in Table 9.8 is also a general effect. It is a marginal and still unresolved failure of the theoretical model for very large brands, not anything specific either to the present data or to Corn Flakes as such.

Despite the apparent complexity of the initial data, the main result is a simple one. The number of people who buy different pairs of brands can be represented, to a first degree of approximation, by a model which merely

depends on the penetration levels of the brands and which has only a single numerical coefficient, D. The reason for this simple kind of result is an important one, and this we now discuss.

9.4 The Other Variables

Various examples in this book have shown how it is possible to establish a simple relationship between a pair of variables whilst ignoring other variables (e.g. the height/weight equation holds despite variations in age, sex, race, etc.). But in the last section this looked more blatant. It seems wrong to try to build a model for the number of consumers who buy brands X and Y while ignoring variables like pack-size, amount bought, price paid, retail outlet, household size, income, the weather, the amount of advertising, and whether they buy yet ***&her*** brands.

The prime justification for leaving out all these variables is simply that the model ***works:*** the analyses lead to simple results which generalise.

But now we can also examine separately these other excluded aspects For example, the preceding analysis ignored how *often* duplicated buyers of X and Y bought each brand. Table 9.9 shows there is little variation in the number of times consumers bought a brand, regardless of which other brand they ***also*** bought in the analysis-period. Thus consumers of Corn Flakes bought them on average 6 times in the 24 weeks, irrespective of whether they also bought Weetabix or Shredded Wheat, and consumers of Puffed Wheat bought that on average about 2 or 3 times.

TABLE 9.9 The Average Number of Purchases by Duplicated Buyers

24 Weeks	The Average Numbers of Purchases of				
	Corn Flakes	Weet-abix	Shred. Wheat	Sugar Puffs	Puffed Wheat
by buyers of					
Corn Flakes	(6)	6	5	4	3
Weetabix	6	(6)	4	3	2
Shredded Wheat	6	6	(4)	4	2
Sugar Puffs	6	6	4	(3)	3
Puffed Wheat	5	5	3	3	(2)
Average	6	6	4	3	2

This kind of result also generalises widely. It therefore illustrates how the duplication law is not the only aspect of the system that follows a simple pattern. It also explains why the average number of purchases could be ignored in the analysis in the last section. The figures in each column of Table 9.9 are more or less constant and therefore cannot affect the other variables!

Similarly, Table 9.10 explains why we could also ignore the extent to which duplicated buyers buy *other* brands. Buyers of each brand made a virtually constant number of cereal purchases in the 24-week analysis-period. Corn Flakes buyers made about 14 purchases, so did Weetabix buyers, and so on.

TABLE *9.10* The Average Number of Purchases of ANY of the Brands, by Buyers of Each Brand of Breakfast Cereals

24 Weeks	By buyers of					
	Corn Flakes	Weet-abix	Shred. Wheat	Sugar Puffs	Puffed Wheat	Ave - rage
The average number of purchases of ANY of the brands	13	14	15	13	15	14

These results illustrate what appears to be the crucial step in analysing multivariate data: to try to split the data into components that can be handled separately, each giving a simple and generalisable result.

9.5 **A Breakdown in Generalisation**

In many cases however, an initially promising result does not generalise. Returning to the children's height/weight example, if we take ***age*** into account as a third variable and look at data for Birmingham boys in 1947, we have three paired equations:

$$\log w = .02h + .76, \quad \text{between height and weight,}$$
$$\log w = .04A + 1.43, \quad \text{between weight and age,}$$
$$h = 2.0A + 33.7, \quad \text{between height and age.}$$

But if we analyse data for British boys in 1880 (see Exercises 8F and 8G), we get different results for weight with age and height with age:

$$\log w = .04A + 1.37,$$
$$h = 2.0A + 30.3.$$

Therefore the boys in 1880 were consistently lighter and shorter than those in 1947, by about .06 log lbs and 3 inches. Since the height/weight relationship remains the same for both groups, the boys still had the same "shape": those in 1880 were simply smaller by the equivalent of about $1\frac{1}{2}$ years' growth.

In the system of three different "paired" relationships only the height/weight equation therefore generalises. The power of this equation is, however, increased by the fact that it now also holds for boys who were markedly smaller at any given age, i.e. by the *failure* of the other equations to generalise.

9.6 **Relationships in More Than Two Variables**

Often similar relationships differ only in one of their coefficients. Then a third variable can be introduced into the relationship to try to account for this variation.

This might be the case with the height/age and weight/age equations, but no work on this problem has yet been done, Another example earlier in Chapter 6 was that the intercept-coefficient of the *height/weight* equation varied somewhat with the apparent nutritional level of the children and therefore a three-variable relationship of the form $\log w = .02h + n$ might result; but here also no analysis on a suitable quantitative measure n of nutrition has yet been carried out.

In contrast, the behaviour of gases provides a well-developed example of a third variable entering into a two-variable relationship. Boyle's Law $PV = C$ relates the pressure and volume of a given body of gas at a constant temperature.

But when the temperature T varies, it has been found that the coefficient C varies proportionally. This gives a more general relationship in *three* variables known as the *Gas Equation:*

$$PV = RT,$$

where R is constant for any given amount of gas.

The Gas Equation contains Boyle's Law as a special case. Thus when temperature T is constant we have again

$$PV = \text{Constant}.$$

Similarly, the Gas Equation contains *Charles' Law*

$$V = KT.$$

This says that when pressure is constant, the volume of any given body of gas expands at a constant proportion of the degree-rise in temperature.

The Gas Equation also contains a third two-variable law

$$P = LT.$$

This says that when the volume of a given body of gas is constant, the *pressure* changes at a constant proportion of the changing temperatures, where L is another constant.

We therefore have a system of three paired equations which differ from those discussed earlier in this chapter. Here the equation for any one pair of variables holds only when the third variable is constant. In the other cases the additional variables did not enter into the paired relationships at all. For example, with height, weight, and age the relationship log w $= .02h + .76$ held *despite* variations in age. Similarly, in Section 9.2 the relationship between brand usage and any particular attitudinal response (say "Good Value") held regardless of the responses given to other attitudinal variables.

In general, equations with large numbers of variables arise either when the numerical coefficients of a simple equation vary and this variation can be dealt with by relating it in turn to another variable or when the *deviations* from the law can be systematically related to some other variable.

Ultimately, the most general laws of science are freed from arbitrary-looking numerical coefficients by introducing such explanatory variables or by choosing appropriate units of measurement, Thus the Gas Equation $PV = RT$ is only that simple if T is measured as the *absolute* temperature with its zero-point at $-273°C$ (one of the "absolute constants" of physical science). And the equation can become $PV = 2T$, replacing the variable coefficient R by the "absolute" number 2 (or 1.987 calories, to be more exact), when the law is applied to an amount of gas equal to its molecular weight (like 2 grams of hydrogen, 16 grams of oxygen, and so on). Here the choice of units is the equivalent to introducing a further variable, the molecular weight of the gas.

9.7 **Correction Factors**

More variables can also be introduced as correction factors, because empirically based relationships are generally oversimplifications that never fit exactly. This is true of all the examples discussed in this book, including the Gas Laws. These hold only for perfect gases, defined, as already mentioned, as substances for which the Gas Laws hold. *Actual* gases follow these laws at all closely only at low pressures. But, in general, correction factors have to be used.

Various adjustments have been developed to give a better approximation of actual gases, even under relatively high pressure. The best-known, although still only an approximation, is Van der Waal's equation

$$\left(P + \frac{a}{V^2}\right)(V - b) = RT,$$

where a/V^2 provides a correction for the mutual attraction of the gas molecules, and b is a correction for the volume occupied by these molecules.

One often makes do with *ad hoc* corrections because too little is known about many phenomena to model them explicitly. For example, when using

the basic result that children's heights and weights follow log w $= .02h + .76$, the coefficient .76 is adjusted to values like .74 or .72, depending on the type of child being studied. A similar allowance is made for the fact that teenage girls tend to be relatively heavy by about .04 log lbs. Again, the duplication of purchase law, $b_{XY} = Db_X b_Y$, is widely used even though it systematically overstates the duplication level for very popular brands. One simply corrects the estimated value by subtracting a few percentage points. Until a great deal more is known about such deviations it is usually pointless to develop more explicit mathematical models.

9.8 Summary

The analysis of multivariate data can usually start by examining the relationships between *pairs* of variables. The greater number of variables does not necessarily lead to particularly complex results. The extra information generally can provide firmer conclusions than when dealing with only two variables.

If some of the fitted equations do not generalise across different sets of data, additional variables or correction factors have to be introduced to try to account for the differences. This is one way relationships involving more than two variables can be developed.

CHAPTER 9 EXERCISES

Exercise9A. FollowinguponthеAppleTrees

In the Apple-Tree example of Section 9.1, what should one study next?

Discussion.

The eight trees measured for each rootstock came from one parent. Thus the apparent differences might be due to the specific parents and not to the different rootstocks.

Studies of these rootstocks using trees of different parentage are therefore required. This would also show any generalization of the larger deviations of certain rootstocks from the overall pattern.

Exercise9B. Following up on Consumer Attitudes

What should one study next in the market-research example in Section 9.2?

Discussion.

The references cited show that the relationship between the attitudinal responses to a brand and its user level is already well-established for many different attitudes and across many product-fields. Therefore three points

to be studied next might be:

(i) the large occasional exceptions illustrated in Table 9.4
(ii) why the relationship occurs, and
(iii) what determined the numerical value of the coefficient in the relationship.

Exercise9C. TheNatureoftheDuplication-of-PurchaseLaw

Retrace the basic steps and outstanding questions in the development of the duplication of purchase law, $b_{XY} = Db_X b_Y$.

Discussion.

There are three main aspects in the development of any such result: conceptual, empirical and theoretical.

It first had to be decided to analyse consumers' purchasing behaviour in specific time-periods, and to do so by examining separately the number of ***people*** who buy an item at all and the number of ***times*** they buy it. (It has not yet been established whether the equation or some equivalent holds when the data are viewed in different ways, e.g. for pairs of successive purchases by each consumer, or for the brand-shares of a consumer's total purchases.)

Having noted that the equation $b_{XY} = Db_X b_Y$ gave an adequate fit to one set of data, the next step was to establish whether it generalised across different length time-periods, different product-fields, different countries, and so on. The nature and generality of the exceptions to the relationship also had to be established (e.g. the "large-brand" effect).

Finally, theoretical questions remain, such as

(i) How does the relationship relate to ***other*** aspects of buyer behaviour?
(ii) How can it be reformulated to account for a consistent exception, such as the "large-brand" effect?
(iii) What determines the numerical value of ***D,*** the one coefficient in the model?
(iv) How can a better understanding of the result be reached?

Exercise9D. Reformulating the Duplication-of-Purchase Law

What does the equation $b_{XY} = Db_X b_Y$ ***mean?***

Discussion.

The proportion of buyers of Brand X who also buy Brand Y can be expressed as

$$\frac{b_{XY}}{b_X}.$$

Thus if $b_X = .20$ and $b_{XY} = .05$ (i.e. **20%** of all consumers buy X and 5% buy X and Y), then $b_{XY}/b_X = .05/.20 = .25$ or **25%** of buyers of X also buy Y.

Since $b_{XY} = Db_X b_Y$, ***we*** have that

$$\frac{b_{XY}}{b_X} = Db_Y.$$

This reformulation says the proportion of buyers of X who also buy Y depends only on b_Y, the penetration of Y, and not on brand X as such. The duplication-of-purchase law therefore says that

> "Consumers of Brand X are D times as likely to buy Brand Y as the whole population, where D is approximately the same for all pairs of brands in a product-group".

Table 9.11 shows this reformulation of the duplication-law in arithmetical form for the data in Section 9.3.

TABLE 9.11 The Percentage of Buyers of One Brand who Also Buy bother Brand

24 Weeks		Who also Bought: Corn Flakes	Weet-abix	Shred. Wheat	Sugar Puffs	Puffed Wheat
% of Buyers of						
Corn Flakes	100%	-	57	41	31	18
Weetabix	100%	71	-	45	35	20
Shredded Wheat	100%	76	67	-	33	27
Sugar Puffs	100%	73	65	42		31
Puffed Wheat	100%	69	61	56	50	
Average	100%	72	63	46	37	24

If the law held *exactly*, the percentages in each column would be identical. The table in fact shows that Corn Flakes were bought by *about* 72% of the consumers of any of the other brands, Weetabix by *about* 63% of the consumers of any other brands, and so on. On the whole, these tendencies to buy one brand do not depend on which other brand is also considered, the deviations are only a few percentage points. With such a simple pattern any exceptions also stand out clearly, like the relatively high duplication between the two Quaker Oats brands, Puffed Wheat and Sugar Puffs. (The message of the duplication-law is that while such groupings or "clusterings" of particular brands might be expected to occur generally, they are in fact the exception.)

TABLE 9.11a The Average 24-week Duplications and Penetration Levels of Each Brand

(D = 48/37 = 1.3)

24 Weeks	Brand: Co. FL.	Weet.	Sh. Wh.	Su. Pu.	Pu. Wh.	Av.
Av. Duplication	72	63	46	37	24	48
1.3 x Penetration	79	64	43	31	21	48
Penetration %	61	49	33	24	16	37

Table 9.1 la shows the second part of the relationship, that the duplication level is proportional to the penetration of each brand. Thus with $D = 1.3$, if 50 % of the population buy a brand, then about 1.3 x 50 = 65 % of the buyers of any other brand should also buy it. The tendency for the theoretical law to overstate somewhat the observed result for very large brands (like Corn Flakes here) stands out clearly against the general pattern. (The theoretical understatement for the two smaller brands reflects the "Quaker Oats" cluster already mentioned above.)

Exercise9E. EmpiricalVariationsin *D*

How can one further investigate the nature of the duplication-coefficient ***D?***

Discussion.

At this early stage some purely empirical "looking" should be rewarding. Thus when establishing the general validity of the duplication law, values of D must have been calculated under many different conditions. Can any patterns be seen in these results?

The references already cited report a general pattern when analysing duplicated purchases for a given set of brands in different length time-periods. It is generally found that, for a particular product-field, the D-values increase from almost zero to more than 1, as the first line of Table 9.12 illustrates (Ehrenberg, 1972). In a short time-period such as a week buyers of X are *less* likely to buy Y than the whole population. Buying of one brand inhibits buying the other. But in a year, buyers of X are *more* likely to buy Y than the whole population.

TABLE 9.12 The Duplication-Coefficients for Brands and Varieties in Different Length Time-Periods - An Illustrative Example

	Analysis-Period, in Weeks				
	1	4	12	24	48
D for Brands	.3	.5	1.0	1.2	1.4
D for Varieties	2.4	1.9	1.5	1.4	1.4

The opposite pattern is reported for duplicated purchases between different *varieties* of a product (e.g. different flavours), as shown in the second line in Table 9.12. In a week, buyers of one flavour are far more likely to buy another flavour too. But in the longer time-periods, this tendency decreases.

These results reflect that in a week, i.e. usually on a single shopping-trip, people seldom buy two more or less identical brands (low ***D).*** But they may well do so on different purchase occasions spread over a longer time-period (high ***D).*** In contrast, purchasers in one week contain a relatively high proportion of frequent consumers of the product, who may well buy a number of different flavours on the same shopping-trip (high ***D).*** In longer time-periods, lighter buyers also show up, and they tend to buy *fewer* flavours (a relatively lower ***D).***

The variations in the value of D for different length time-periods therefore distinguish between items which are substitutable (i.e. brands of similar or identical product-formulation) and items which are complementary (i.e. different varieties of a product, such as different flavours). This is an example of how additional variables (here length of time-period and type of product) can relate to a coefficient in a given relationship, as was discussed more generally in Section 9.6.

Exercise 9F. **TheEffectsofAdvertising**

Why is there no need to allow for *advertising* in either the attitudinal or the duplication-of-purchase models?

Discussion.

The results described essentially apply when there are no very marked trends or fluctuations in the sales levels of the different brands. (This is what occurs in most markets most of the time.) It follows that advertising and other marketing variables are then having no ***positive*** effect on sales. (Their roles are mainly defensive, to maintain the status quo.)

Exercise9G. TheRolesofOtherVariables

Boyle's Law, $PV = C$, develops into the more general Gas Equation, $PV = RT$, if temperature T varies, but Age A does not enter into the height/weight relationship, log w $= .02h + .76$. Why not?

Discussion.

Any empirical law holds only when certain other variables remain the same. Yet other factors may still vary. For example, log w $= .02h + .76$ holds for children under "normal" conditions : standing upright, effectively weighed without clothing, etc. However, it is found that age can vary without affecting the relationship, and so can the weather! Neither age nor the weather can therefore enter into the relationship.

Boyle's Law holds when there is a fixed amount of gas, no chemical reaction, and effectively constant temperature. If these factors vary, the law breaks down. The effect of varying temperatures is, however, rather simple, namely that the coefficient C varies in proportion. Hence we get $PV = RT$.

Exercise 9H. TheValueofMoreData

In a recent study of certain leading manufacturers in a number of countries, each company was assessed on six variables A to F (e.g. its Research, its Product-quality, etc.). Relating the percentage scores of the six variables to a measure of overall standing S for each company gave the following working-solutions for two of the countries:

U.K.: $A = .76S - 3$, $B = .58S$, $C = .25S + 5$, $D = .31S + 1$,
Germany: $A = .56S - 1$, $B = .51S - 1$, $C = .44S$, $D = .07S + 3$,
etc.

The fit was generally within a few percentage points, but there were also some very large exceptions (10 to 20 points).

The working-solutions vary mostly from one country to the other. How can this be further analysed?

Discussion.

We need to check that the occasional very large deviations have not unduly affected the working-solutions.

In the present instance it was found that differences between the equations for the U.K. and Germany always appeared to be caused by an exceptional value having affected the fitting process. Further analysis showed that when the *same* equation was fitted to different countries, the fit was almost as good as that of the original working-solutions.

This might not have been seen with just two countries. It became more apparent when a larger number of countries and all six variables were analysed, with the following results:

U.K.: $A = .6S,\ B = .5S,\ C = .4S,\ D = .3S,$
Germany: $A = .6S,\ B = .5S,\ C = .4S,\ D = .3S,$
France: $A = .6S,\ B = .5S,\ C = .4S,\ D = .3S,$ etc.
etc.: $A = .6S,\ B = .5S,\ C = .4S,\ D = .3S.$

Against these general norms, deviations for particular companies, countries, or variables stand out clearly and can be further investigated.

CHAPTER 10

The Emergence of Theory

The main function of theory is to integrate a wide range of results into a single conceptual framework or model. Thus the discussion in Chapter 9 had an increasingly theoretical orientation because we were interrelating the results for many variables.

Theoretical considerations tend to dominate most analyses, except at the very earliest stages of reducing a new kind of data to summary figures. While a detailed discussion of theory is beyond the intended scope of this book, in this chapter we illustrate how a higher level theoretical result can emerge.

10.1 Different Levels of Theory

Theory operates at many different levels. For example, when we fitted the height/weight relationship, $\log w = .02h + .76$, we were dealing with a low-level theoretical abstraction.

We moved up a step in our "theoretical" approach to data-handling when we used the equation to analyse additional height and weight data. But there was still nothing very advanced about this.

We can move to a still higher level of theory by deriving the height/weight relationship theoretically instead of empirically. For example, from the 1947 data on the Birmingham boys in Chapter 9 we had two equations:

$$h = 2.0A + 33.7, \quad \text{between height and age,}$$

$$\log w = .04A + 1.43, \quad \text{between weight and age.}$$

We derived these two equations empirically, from the data. If we now eliminate the common variable A from both (by multiplying the first equation by .02 and then subtracting the result from the second equation), we again arrive at the height/weight relationship $\log w = .02h + .76$, as we have already been doing in the last chapter.

This derivation is theoretical because we did not here *directly* relate empirical data on heights to data on weights. We did it indirectly or

"theoretically" instead. Theoretical work like this is usually simpler than direct numerical analysis of data, as long as one is equipped with the required mathematical expertise.

The Role of Hypotheses

Hypotheses are theoretical formulations or assertions which are not yet known to be true, i.e. unproven suppositions. Their role is to suggest either new facts to collect or new analyses to carry out. Thus the hypotheses can be tested.

It is mainly through such speculation and hypothesising that one discovers new truths. For example, when we were fitting the height/weight relationship in Chapter 8 we used speculative theory to suggest the hypothesis of a cube-root transformation of weight. Without the "dimensional" theorising that weight tends to be proportional to volume, and that volume is proportional to the cube of a linear dimension like height, we would not have thought of trying a cube-root transformation. It simply is not the kind of idea that springs to mind naturally. In fact, most scientific relationships have forms that are far too complex to be based on mere common sense.

However, hypothesising can easily suggest things which are not true. From the height and age equation $h = 2.0A + 33.7$ for the Birmingham boys in 1947, and the *weight* and age equation $\log w = .04A + 1.37$ for the British boys in 1880, the same kind of theoretical elimination of the common age-variable as above would lead to the hypothesis that height and weight should be related as $\log w = .02h + .70$.

This hypothesis now has to be tested against facts, i.e. data on both height and weight for the same boys. It is then found that it does not hold either for the 1947 or the 1880 boys. Instead of being .70, the intercept-coefficient for these data is .76. The explanation is that the height/age and weight/age relationships in 1947 and 1880 differed (see Section 9.5), and therefore we cannot mix one equation from one set of data and one from the other.

Formulating and testing such a speculative hypothesis would, however, have been a proper and useful thing to do. Often we have only limited data from any particular study, e.g. height and age for 1947, and weight and age for 1880. Scientific progress largely consists of deducing new hypotheses from such incomplete information and then eliminating those which are not true by checking against new facts.

But theoretical arguments and assumptions based on insight or hunch must be sharply distinguished from *validated* theory. The one reflects what we think, the other what we know. It is the failure to make the distinction that has often given "theory" a bad name.

The Main Role of Theory

The function of speculative theory and hypotheses in discovering a particular scientific result is not, however, of lasting importance, except to the history of science. The principal role of successful theory is to link different kinds of *known* results. Thus, the cube-root relationship links the height and weight data with our more general experience of the shape and density of bodies. The link is approximate and suggestive rather than exact and compelling, since children are not rectangular nor perhaps of absolutely constant density as they grow. But this does not detract from the broad conceptual simplification achieved.

Theory is most powerful when it works at a detailed quantitative level and is empirically well-based. In the rest of this chapter we shall illustrate this integrative and explanatory power of more advanced theory with an example.

10.2 A Trend in Purchasing Frequencies

The problem to be tackled is that of finding a model for a series of empirical results which are individually simple to describe but which as a body are complex. Mere inspection of the data or direct "curve-fitting" as practised in the preceding chapters is unlikely to produce useful results.

Instead, the required result can be deduced from *other* findings. To illustrate, we recall the discussion of the five different brands of breakfast cereals in Chapter 9. Each brand was bought at more or less the same average rate per buyer in a 4-week period, as is shown again in Table 10.1. These purchase rates, which are usually called w, vary only by about ± 0.2 from the average 1.7. This variation is small compared with the differences in market-shares of the five brands, which range from 38 % down to 5 %. (These shares arise from multiplying the penetrations from Table 9.7 and the purchase frequencies from Table 9.9 and then percentaging.)

TABLE 10.1 **4-Weekly** Purchase Frequencies

4 WEEKS	Corn Flakes	Weet-abix	Shred. Wheat	Sugar Puffs	Puffed Wheat	Ave-rage
% Market-Share	38	31	16	10	5	(100)*
Purchases of the stated brand per buyer of the brand in 4 weeks	1.8	2.0	1.7	1.5	1.6	1.7

* Total sales of the 5 brands

But, in a longer period of 24 weeks, buyers of the different brands bought them at markedly different rates, the values of w ranging from 6 down to 2,

TABLE 10.2 24-Weekly Purchase Frequencies of Different Brands of Breakfast Cereals

24 WEEKS	Corn Flakes	Weet-abix	Shred. Wheat	Sugar Puffs	Puffed Wheat	Ave-rage
% Market-Share	38	31	16	10	5	(100)*
Purchases of the stated brand per buyer of the brand in 24 weeks	6	6	4	3	2	4

* Total sales of the 5 brands

as shown in Table 10.2. These variations seem to be related to the brands' market-shares.

We therefore have a highly discrepant situation : a trend in w in one case and approximately constant values of w in the other. This complex pattern generalises for a wide range of other product-fields, food and non-food, in the U.K. and U.S.A., etc. The data show that the longer the analysis-period, the stronger the trend between average purchase frequency w and market-share (with hindsight we can even discern a very weak trend with market-share in Table 10.1, and this too generalises for other products in short time-periods).

It seems that we need a complex model involving three variables :

(i) the average purchase frequency w of each brand,
(ii) its market-share,
(iii) the length of the analysis-period.

10.3 A Theoretical Model

A simpler answer is reached by introducing another variable altogether, namely the *penetration* of each brand, which is called b. This we have already come across in the last chapter. It is the proportion of the population who buy the brand at least once in the time-period. It varies both with the brand's market-share (the higher the brand's share of total sales, the more people buy it) and with the length of the analysis-period (more people buy a brand in two weeks than in one week, but the value of b does not simply double, since some people buy in *both* weeks). The relationship between penetration b and market-share therefore *varies with the length of the analysis-period.* This is the kind of pattern needed to model the varying trends in w we saw in the purchase frequencies.

The specific form of the theoretical model follows mathematically from three results which have already been described towards the end of Chapter 9. There we had that for any two brands X and Y :

(A) Consumers of Brand X buy the total product-class at about the same average rate as do consumers of Brand Y (Table 9.10).

(B) The proportion of the population who buy both X and Y in the same analysis-period is given by the equation $b_{XY} \doteq D b_X b_Y$ (Table 9.8).

(C) A duplicated consumer buys a brand at about the same rate of purchase as all its other consumers (Table 9.9).

We now explore what follows mathematically from these three results. For simplicity we do so in terms of only three brands, X, Y and Z (the mathematical argument readily extends to more than three brands).

We start with the average rate of purchase of the total product-class by the buyers of Brand X. The number of buyers is Nb_X, i.e. N, the total number of potential consumers in the population, times b_X, the penetration of X. Their purchases of the total product-class are what they buy of Brand X, of Y, and of Z. This can be expressed as

$Nb_X w_X$ = the number of buyers of X times how often on average they buy X (i.e. w_X),

$Nb_{XY} w_{Y \cdot X}$ = the number of buyers of X who also buy Y (Nb_{XY}), times how often on average they buy Y (where $w_{Y \cdot X}$ denotes the average rate of purchase of Y by those buyers of X who also buy Y),

$Nb_{XZ} w_{Z \cdot X}$ = the number of buyers of X who also buy Z, times how often they buy Z.

Therefore we have that the average purchases of the total product-class by buyers of Brand X is

$$\frac{Nb_X w_X + Nb_{XY} w_{Y \cdot X} + Nb_{XZ} w_{Z \cdot X}}{Nb_X}$$

Now from Item (A) above we know empirically that the average rate of purchase of the total product group is the same for buyers of Brand X as for buyers of Brand Y. Therefore

$$\frac{Nb_X w_X + Nb_{XY} w_{Y \cdot X} + Nb_{XZ} w_{Z \cdot X}}{Nb_X} = \frac{Nb_Y w_Y + Nb_{YX} w_{X \cdot Y} + Nb_{YZ} w_{Z \cdot Y}}{Nb_Y}.$$

But from Item (B) above we also know that

$$b_{XY} = b_X b_Y \text{ and } b_{XZ} = b_X b_Z, \text{ etc.}$$

if we take the simple case where the coefficient $D = 1$. (The more general case where D is not equal to 1 is discussed in Exercise 10D.) From Item (C) above we know empirically that

$$w_{Y \cdot X} = w_Y, \text{ and } w_{Z \cdot X} = w_Z, \text{ etc.}$$

If we substitute these values in the complex equation above and cancel through by N, we get

$$\frac{b_X w_X + b_X b_Y w_Y + b_X b_Z w_Z}{b_X} = \frac{b_Y w_Y + b_Y b_X w_X + b_Y b_Z w_Z}{b_Y}.$$

Now if we cancel through by b_X on the left and by b_Y on the right, and eliminate the common term $b_Z w_Z$ from both sides, we get

$$w_X + b_Y w_Y = w_Y + b_X w_X,$$

or, collecting terms in w_X and w_Y,

$$w_X(1 - b_X) = w_Y(1 - b_Y).$$

This is the result we require-an equation which relates how the average rates of purchase w_X and w_Y of Brands X and Y vary to how *another* variable, the penetrations b_X and b_Y, varies from brand to brand.

We can write the equation as $w_X(1-b_X)=c$, a constant, or dropping the suffix,

$$w(1 - b) = c$$

for the different brands in the given length of analysis-period. In other words, w varies with *b,* namely as $w = c/(1-b)$, a very simple result.

How it Works

It is tempting to try to read some direct "meaning" into this theoretical result. But like most laws of science, it merely describes and interrelates observed phenomena and has no obvious "commonsense" meaning. The crucial feature of such a theoretical result is how it works and how it links up with other findings.

Table 10.3 gives the values of $w(1-b)$ for the five brands in our example in the 24-week and the 4-week periods that have been analysed. The values of $w(1-b)$ are not *precisely* constant in each time-period, but the variation is relatively small, at about $\pm 10\%$. More important is the fact that the 24-week figures no longer show a trend with market-share. The "correction factor" $(1-b)$ has therefore accounted for the trend in w.

TABLE *10.3* *The* Values of w(1-b) for the Five Brands

	Corn Flakes	Weet-abix	Shred. Wheat	Sugar Puffs	Puffed Wheat	Ave-rage
% Market-Share	38	31	16	10	5	(100)
24 weeks	2.2	2.9	2.9	2.5	2.2	2.5
4 weeks	1.2	1.5	1.5	1.4	1.5	1.4

This general no-trend pattern for $w(1-b)$ has also been found in many other product-fields. (The rather low values for Corn Flakes here do *not* represent a general feature for market leaders.)

The formula $w(1-b)$ works in a relatively complex way. It eliminates the strong trend in w in the 24-week period, but does not introduce a contrary

pattern for the data in the 4-week period. The formula achieves this because of the mathematical nature of the quantity $(1-b)$.

In long time-periods, the penetration of leading brands is relatively high, say .4 or .6: i.e. 40% or 60% of the population buy the brands at least once. As the values of b increase the values of $(1-b)$ decrease at a far greater rate, as Table 10.4 shows. When $b=.8$, $(1-b)$ is twice as large as when $b=.9$ and 20 times as large as when $b=.99$. Multiplying by $(1-b)$ therefore has a great effect when b is large.

TABLE 10.4 Values of (l-b) for Large b

b	.8	.9	.99
l-b	.2	.1	.01

In contrast, for short time-periods, the penetrations of most brands are relatively low, say .2 or .1: i.e. only 10% or 20%. Then the values of $(1-b)$ are all close to unity, as shown in Table 10.4a. Multiplying by $(1-b)$ in these cases has relatively little differential effect. The difference in $(1-b)$ for $b=.01$ or $b=.20$ is only about 25%, which is small compared either with the difference in b itself (a factor of 20, i.e. 2000%) or with the differences in sales levels or market-shares.

TABLE 10.4a Values of (1-b) for Small b

b	.01	.1	.2
l-b	.99	.9	.8

The same factor $(1-b)$ can therefore do two things. It can reflect differences in w which occur in relatively long time-periods because b is relatively high then. And it can reflect the ***absence*** of large differences in w which occurs in relatively short time-periods, when b is relatively low. The length of the analysis period, one of the variables clearly at work as noted earlier, is therefore taken into account indirectly.

The theoretical importance of the model $w(1-b)=$ constant is that it interrelates a number of separate empirical results into a single theory. Instead of having an isolated formula for the value of the average purchase frequency of a brand, we have a model that relates this to

(i) consumers' known brand duplication of purchases and

(ii) their total rates of product usage.

Given that these phenomena exist as they do, the formula $w(1-b)$ must follow. It is through this integrative function that more advanced theory is built.

Three stages are normally involved in such developments. First a simple regularity is noted, e.g. that w in a short time-period is more or less *constant*

(largely ignoring such small variations and even the possibility of a trend as might seem to exist in tables such as 10.1). Secondly, rather different results arise, e.g. that in longer time-periods w *varies* with the market-share. Finally, both types of results can be accounted for by the same model and this follows as a consequence of using *other* findings (e.g. the brand-switching law $b_{XY} = Db_X b_Y$, etc.).

It would be difficult to improve on the words used by Sir Cyril Hinshelwood (1967), uniquely President of the Royal Society and the Classical Association in the same year, to describe this sequence of stages. The first stage he described as

> "Gross over-simplification, reflecting partly the need for practical views and even more a too-enthusiastic aspiration for the elegance of form." *(Constants* indeed!)
>
> In the second stage, "the symmetry of the hypothetical system is distorted and the neatness marred as recalcitrant facts increasingly rebel against uniformity".
>
> In the third stage? "if and when this is obtained, a new order emerges, more intricately contrived, less obvious and with its parts more subtly interwoven, since it is of nature's and not of man's conception".

10.4 Summary

Theory operates at different levels. *Any* abstraction from the observed facts is theoretical, at least at a low level. But as different results interrelate, more advanced theory and understanding begin to grow.

Theoretical arguments can be purely hypothetical and these must be sharply distinguished from *validated* theory. The ultimate role of valid theory is to integrate and interpret results which are well-founded in fact.

CHAPTER 10 EXERCISES

Exercise 10A. IsitPractical?

Of what practical use is a theoretical result like $w(1-b) =$ constant?

Discussion.

Two practical applications are in assessing sales targets for new and established brands. For simplicity's sake we shall consider relatively short time-periods. The penetrations b of different brands are then all low so that the factor $(1-b)$ hardly varies, being close to 1 for all brands. The relationship therefore simplifies to w $\doteqdot$ constant. This is illustrated in Table 10.5 for breakfast cereals in a 4-week period.

TABLE 10.5 Purchase Frequencies over 4 Weeks
(Repeated from Table 10.1)

4 WEEKS	Corn Flakes	Weet-abix	Shred. Wheat	Sugar Puffs	Puffed Wheat	Ave-rage
% Market-Share	38	31	16	10	5	(100)*
Purchases of the stated brand per buyer of the brand in 4 weeks	1.8	2.0	1.7	1.5	1.6	1.7

* Total sales of the 5 brands

A New *Brand.* Suppose that a new brand of breakfast cereal is to be launched for which a rate of sales of about 2 purchases per housewife per year has been set (for when the sales "settle down" six months or so after the launch). How can this target be achieved?

Two purchases per housewife per year is about .15 purchases per housewife per 4 weeks, or 15 purchases per *100* housewives, expressed on a per hundred basis to keep the arithmetic simple. A 4-weekly sales target of 15 purchases per 100 housewives could be achieved in various ways, e.g. 1% buying 15 times each, 5% buying three times each, 10 % buying on average 1.5 times each, or 15 % buying just once each in the 4 weeks.

But Table 10.5 shows that over 4 weeks the average purchase frequency w of established brands of breakfast cereal is about 1.7. It follows that about 9 % of the population will have to buy the new brand at the predictable average rate of 1.7 purchases each to reach the target of 15 per 100 housewives. (The variation in the w's in Table IO.5 implies some variability in the required penetration, from 8 % to 10 %.)

We are not here predicting *sales*, the $64,000 question. The prediction is only that, whatever happens, the new brand will tend to be bought 1.7 times in 4 weeks by its buyers. What cannot be predicted is how many buyers there will be, the figure of 9% was the penetration which is required to achieve the given sales *target*.

Instead of having two unknowns *b* and *w*, we have a firm prediction for w (worth $64!) and a *target* for *b*. We now know what to plan for, a penetration of about 9%, and we have a yard-stick for assessing test-market and launch results.

The prediction for w is firm unless there is a specific reason to expect the purchase frequency for this particular new brand to be radically different from that of the others in the product-class. The new brand would have to differ more from the existing cereal brands than they do from each other.

An Established Brand. To increase the sales of an *established* brand, like Shredded Wheat, one could in principle aim to increase either *b*, the number of people buying it, or w, the rate at which existing buyers buy it, or both *b* and w.

But we now know that the 4-weekly rate of sales of Shredded Wheat cannot be doubled by simply doubling its average purchase rate from the value 1.7 in Table 10.5 to 3.4, since nothing like the latter rate of purchasing has ever been observed for any cereal brand. It follows that

if sales *are* to increase. It has to be through an increase in the number of buyers in 4 weeks. The alternatives have been ruled out by the empirical constraints of consumer behaviour.

Exercise 10B. The Units of Measurement

In the last exercise, sales were equated to the number of buyers multiplied by their rate of purchase, ignoring *how much* they bought per purchase occasion (e.g. whether large or small packs, and how many at a time). How can this be?

When discussing consumers' purchasing behaviour earlier, we sometimes spoke of numbers of purchases and sometimes of numbers of packs bought. Which is it?

Discussion.

The choice of units is often crucial in developing simple and generalisabie theory. Thus the various laws of consumer behaviour mentioned generally hold only if the data are expressed as *purchase occasions.* (This allows purchases of different pack-sizes, for example, to be dealt with by *ignoring* the pack-size.) The primary justification of this approach is that it works, in giving simple and generalisable results.

In more detail, the total level of purchasing or of sales in a given period can be decomposed as follows:

Sales = (Number in population) x (Proportion buying at all)
x (Average number of purchase occasions per buyer)
x (Av. number of packs bought per purchase occasion)
x (Av. size or price of pack).

Here the number in the population is fixed. The size or price of a pack is usually more or less fixed or known. The average number of packs bought per purchase occasion is approximately constant (e.g. for different brands, for light or heavy buyers, etc.). This is an *empirical* finding (Ehrenberg, 1972). So is the fact that w is approximately constant in a 4-week period. It follows that the only *variable* component of sales is b, the penetration.

The data in the breakfast cereal example came in the form of number of *pucks* bought, so some of the results were referred to like that. But the theory is formulated on the basis of purchase occasions (two packs bought at the same time are treated as *one* occasion). The conflict is resolved because on average just over 1 pack of cereal is bought per purchase, so the distinction between units bought and purchase occasion is in this case numerically trivial-the results hold whichever unit is chosen. In contrast, for a product like petrol, the average purchase is usually just over 3 gallons in Great Britain (and more in the United States say); for this product, theoretical results like $w(1-b)$ operate *only* when the purchase occasions are used as the analysis unit.

Exercise 10C. The Quantities b and w in Different Length Time-Periods

Tables 10.1 and 10.2 show that the average consumer purchased Corn Flakes 1.8 times in 4 weeks and 6 times in 24 weeks (or 5.7 to two digits). Why not 10.8 purchases (6 x 1.8) in a period six times as long?

Discussion.

If sales are steady, the total number of purchases made increases in proportion to the length of time-period. Thus the 491 households on which our data here are based (Charlton *et al.*, 1972) made 281 purchases of Corn Flakes in the 4-week period and 1684 purchases in 24 weeks, which is virtually six times 281. (These figures are equivalent to 57 and 340 purchases per *100* households in 4 and 24 weeks.)

But the average purchase frequency w refers to the frequency *per buyer* in the period. Both w and the number of buyers change with the length of the analysis-period. In 4 weeks, 157 of the 491 households bought Corn Flakes, a penetration of 157/491 or $b = .32$ or 32%. These households made an average of 281/157 purchases in 4 weeks or $w = 1.8$.

In 24 weeks, 298 households bought Corn Flakes, a penetration of $298/491 = .61$ or 61%. They made 1684 purchases, so $w = 1684/298 = 5.7$.

As the length of the analysis period increases, the proportional increase in sales is made up of *less than* proportional increases in both b and w, so that

$$4 \text{ weeks}: \quad 32 \times 1.7 = 54,$$

$$24 \text{ weeks}: \quad 61 \times 5.7 = 348.$$

(The differences from the observed sales levels of 57 and 340 are due to rounding off.)

Exercise 10D. Oversimplification in the Theory

In the theoretical deduction of the model $w_X(1-b_X) = w_Y(1-b_Y)$ it was assumed that the coefficient D in the equation $b_{XY} = Db_Xb_Y$ was equal to 1. Does it matter that this is usually not so?

Discussion.

In addition to D generally not being equal to 1, Item (C) used in the deduction is also generally not quite true. The average purchase frequency $w_{X \cdot Y}$ of Brand X per duplicated buyer of X and Y is usually lower than w_X: i.e. $w_{X \cdot Y} \doteq Cw_X$, where C is a constant. The value of C is often about 0.8 (but in the example here it happened to be 1.0).

The more accurate equations $b_{XY} = Db_Xb_Y$ and $w_{X \cdot Y} = Cw_X$ can be used in a theoretical argument exactly like that of Section 10.3. This gives the result

$$w_X(1 - CDb_X) = w_Y(1 - CDb_Y).$$

This is similar in form to $w_X(1-b_X) = w_Y(1-b_Y)$. The results are almost identical when D is about 1.3 and C is about 0.8, which is often the case. More generally, the two forms tend to give similar numerical results.

Even though the more exact argument is mathematically soluble, at this stage the formulation $w_X(1-b_X) = w_Y(1-b_Y)$ is preferred because it is much simpler. For example, it can be used without first establishing the specific values of C and D for any particular situation.

Using *exact* results in theoretical work commonly leads to more complex mathematics than used here, which is often literally impossible to solve. Simplifying assumptions are therefore commonly introduced, e.g. the notion that the buying of X and Y is simply *independent* rather than

linked by $b_{XY} = Db_X b_Y$. It follows that theoretical deductions are often not altogether true, because the oversimplification was too drastic. This is why theoretical deductions always have to be checked against the facts.

Exercise 10E. By Fits and Starts

Medawar (1952) has noted that many scientific papers do not describe how the results were obtained (some facts first and then a "theory" to explain them) but are written with hindsight in an unnaturally "logical" form, as if the theory had come first. How was the result $w(1-b) =$ constant *really* obtained?

Discussion.

First came the empirical result that $w_X \doteqdot w_Y$ in relatively short time-periods. Next, Dr. C. Chatfield noted that if the product-rates of buying were constant (item A in Section 10.3) and if one assumed that buying of one brand was independent of buying another (essentially items (B) and (C)), then it could not be true that w_X equals w_Y. Instead, the w's should show a trend.

However, by this time it had been established empirically that, especially in longer time-periods, w did in fact show some kind of trend with market-share. This led Mr. G. J. Goodhardt to invert Chatfield's earlier argument giving the result $w(1-b) =$ constant along the lines of Section 10.3. The formula accounted for all the empirical facts, both the trend of w in longer time-periods and its approximate constancy in shorter time-periods.

Exercise 10F. Obvious After the Event?

What does the result $w(1-b) =$ constant mean'?

Discussion.

Table 10.2 shows that consumers of Corn Flakes bought them 6 times in 24 weeks and consumers of Puffed Wheat bought it only twice.

One explanation of such a difference in purchase rates might be that the Puffed Wheat packs are bigger. But this is in effect not so. The same $w(1-b)$ type of result holds for each separate pack-size and also in product-fields such as petrol where different brands are virtually identical in performance and packaging.

Another possible explanation might be that the sort of households that buy Corn Flakes need more breakfast cereal than Puffed Wheat consumers, and that the households satisfy their total needs with the brand they have chosen. But this is not so either. The *total* purchases of breakfast cereals are much higher, about I4 purchases for the average consumer of each brand (Table 9.10). And in any case they are constant across the different brands.

Basically, the meaning of $w(1-b) =$ constant lies in how it describes the data and in the form of its theoretical deduction. If consumers buy different brands more or less independently of each other (i.e. according to items A, B, and C in Section 10.2) then the rates at which they buy individual brands *must* follow something like the relationship $w(1-b) =$ constant.

The temptation to read more into such a law is particularly strong in the present case because not only w_X but also $(1-b_X)$, the other item in

the expression, has a direct meaning of its own: it is the proportion of the population which does *not* buy Brand X in the analysis-period. Why then should the rate of buying per buyer times the proportion of non-buyers be constant for different brands?

The explanation is that the apparent "meaningfulness" of this particular formula is accidental. For example, using the more precise results of Exercise 10D, the expression reads $w_X(1 - CDb_X) = w_Y(1 - CDb_Y)$, and this is sufficiently complex no longer to cry out for any "simple interpretation".

Exercise 10G. Explaining Consumers' Attitudes

Explain the relationship between consumer attitudes and brand usage in Section 9.2, and the occasional large deviations from the relationship.

Discussion.

Since the attitudinal responses are related to the incidence of users, cross-analysis by the "users" and "non-users" of each brand seems a promising next step to explore.

Table 10.6 illustrates typical results. For the attitude "Right Taste", 67% of the users of Brand A say A has it, 62% of the users of Brand B say it about B, etc. In contrast, only 6% of the *non-users* of Brand A say A has the right taste, only 4% of the non-users of Brand B say it about B, and so on.

TABLE 10.6 Percentage of Users and Non-Users of a Brand Holding an Attitude Towards It

	% of the Population using	"Right Taste"		"Convenient"	
		Users of Stated Brand	Non-Users of the Brand	Users of Stated Brand	Non-Users of the Brand
Brand A	50	67	6	19	3
Brand C	30	62	4	55	48
Brand B	10	69	5	17	2
Brand D	5	60	3	17	2

These results generalise. A certain percentage of users of a brand say it has the given attitudinal property, and a much smaller percentage of non-users say so. This holds for many different attitudinal variables and many different brands and product-classes.

Since brands differ in their number of users, a higher proportion of the total population gives a positive attitudinal response for market-leaders than for small brands, the finding in Chapter 9. Thus for Brand A, $(50 \times 67 + 50 \times 6)/100 = 36\%$ of the population should say "Right Taste" about it, and for Brand D only $(5 \times 60 + 95 \times 3)/100 = 6\%$.

The results in Table 10.6 also explain the major exceptions to this general pattern: the results for Brand C and 'Convenient" are quite different. The percentage of non-users of Brand C who regard it as "Convenient" is only fractionally lower than that amongst users (48% and 55%), and both percentages are far higher than for the other brands. This accounts for the exceptionally large proportion of all consumers who regarded Brand C as "Convenient" in Table 9.4.

Such exceptional responses generally occur when a brand differs physically in the relevant respect from the other brands. Thus an indigestion or headache remedy which can be taken as a tablet is more "Convenient" than one which requires water and a glass. Consumers, both users and non-users, notice this and say so. In contrast, when different brands are similar in product-formulation, consumers' attitudinal responses are essentially different ways of saying that they know of or use or like that type of product, as exemplified by the brand.

Although the original undigested data in Table 9.3 in the last chapter looked fairly complex, the results after analysis are beginning to "hang together". For example, it now seems that behaviour here influences attitudes rather than that attitudes cause behaviour. The popular notion of different brands generally differing in their "images" cannot be rooted in any discernible facts. Consumers seem to give the product "image" to the brand they use, i.e. they see it as typical of the product-class. Only when there is a real difference is this reflected in the image, but among both users and non-users. Such conclusions affect our view of the nature of competition among similar products and of the way advertising works in a consumer society (e.g. Ehrenberg, 1974a).