RELATIONSHIPS : LINKS BETWEEN TWO OR MORE VARIABLES

5.1 Introduction

In the previous two topics, we concentrated entirely on distributions and measures of one variable; but in reality, we normally collect data on several items at once. We are interested in links, or relationships, between the different variables (or, sometimes, between variables and attributes).

For example, the fish catch by a local fishery would be affected by many factors, which may include the following:

- the number of people and boats engaged in fishing
- fishing technique used
- weather conditions prevailing
- the surface temperature of the water
- the effect of other fisheries operating nearby.

No doubt many other items could be added to the list. The mathematics of trying to measure the interrelationships of all of these factors would be very complicated. This is referred to as ´multivariate´ analysis, and is beyond the scope of the present course.

We can, however, study the relationship between two variables. Data on two variables are termed bivariate data, and if these are plotted as points on a graph, with one variable on each axis, we have what is known as a scatter diagram. We introduced this briefly in Topic 2. If we look back to Figure 2.1, we see that it demonstrated a relationship between fish catch and number of boats engaged. Similarly, Figure 2.2 showed the relationship between total fish catch and time. In fact in that diagram we drew lines to link up the points on the graph, but we need not have done so. If we had omitted the lines, and plotted only the points, we would have had a scatter diagram of exactly the same type as Figure 2.1.

5.2 Regression

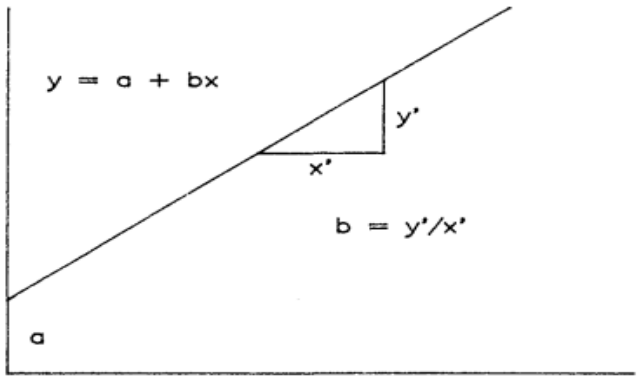
Finding a mathematical formula to describe the relationship

So the purpose in drawing a scatter diagram is to try and get some idea of a simple relationship between two variables. We are not trying to find some mathematical formula that will go through all the points exactly. It is theoretically possible to do this, but the formula would be too complicated to be of practical use. What we would like is some kind of simple formula that ´fits´ or describes the data fairly well. If we can do this, then we have some kind of model that tells us something about the underlying process that produced the data and can help us to make predictions or other decisions. Now with two variables, x and y, the simplest kind of relationship between them is shown on a graph as a straight line. This means that if we increase ´x´ by a constant amount then ´y´ will also increase by a fixed amount.

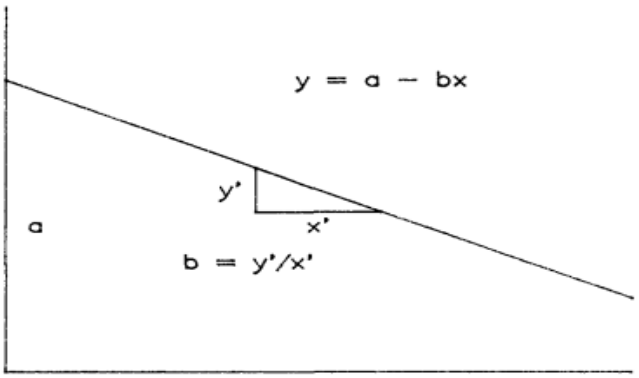
Mathematically, we can represent a straight line by the equation, or formula, $y = a+bx$; a and b are constants where ´a´ represents the point at which the line meets the y-axis, and ´b´ represents the slope of the line. This is shown in Figure 5.1. By changing the value of a and b we change the position of the line on the graph. If the line goes from the bottom

left hand corner to the top right hand corner, then the slope b will be positive. If it goes the opposite way, from the top left hand corner to the bottom right, then the slope will be negative, and b will be a negative number.

FIGURE 5.1 : THE EQUATION OF A STRAIGHT LINE



Positive slope



Negative slope

When we have a scatter diagram, what we want to do is to find a line which best 'fits' the data, that is, which is closest, in some sense, to all the various data points. This means, effectively, to find values for a and b, since it is these two values that define the line. We can undertake this process by eye. Using a scatter graph and a transparent ruler we can move it until it appears to be the best 'fit' to the data, but this is rather unscientific. We have no guarantee that two different people will produce the same line for the same data. Their ideas of the line of 'best fit' may be rather different, and so it will be very difficult to generalise. Instead of using this method then, we use a mathematical technique in which a and b are calculated from the data values (x_i, y_i).

Even when we try to develop a mathematical technique, there are some problems which really arise from the basic situation we are dealing with. In Topic 2, we introduced the concept of independent and dependent variables. We saw that we often had the situation where we were interested

in looking at the way one variable changed as the value of some other variable was altered. So in Figure 2.2, we saw how the level of fish catch changed over time. In this example we say that total catch 'depends' upon time, and therefore total catch, being the dependent variable, must be plotted on the y-axis. In similar fashion, in Figure 2.1, fish catch was the dependent variable in its relationship to number of boats engaged. It is this type of relationship, in which one variable is dependent on another, that we are concerned with when we try to find a line that best fits the data. The purpose of finding the equation of the line, of estimating the values a and b, is that we can then estimate different values of y, given the appropriate values of x.

There are many criteria by which we might define what we mean by 'best fit', but the generally accepted method is the least squares criterion. By this we mean that we will seek to establish the formula which expresses y in terms of x in such a way that the sum of the squares of the differences between the observed values of y, and the values calculated by the formula, is as small as possible.

This technique, of estimating the values of an equation of a line, is known as regression. The line y=a+bx is called the 'regression line' and the values 'a' and 'b' are called the 'regression coefficients'. The equation of the regression line is the formula we shall use to predict the values of y we are likely to get, given certain values of x. We also call this the 'regression of y on x'; y is the dependent variable and x the independent variable. In our first example in the previous paragraph, then, we can talk about the regression of fish catch on time. We want to be able to find out what level of catch we might expect at some future time, by use of a mathematical relationship.

Our data consist of a series of pairs of values, x_i and y_i. We calculate our coefficients, a and b, from these observations. We have:

$$b = \left(\sum_{i=1}^{n} (x_i - \bar{x})\ (y_i - \bar{y}) \right) \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2$$

or, in shorthand form:

$$b = \frac{\Sigma (x - \bar{x})\ (y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

and

$$a = \bar{y} - b\bar{x}$$

We shall look at the expression $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ again in more detail in the next section. We notice that the denominator of b is $\sum_{i=1}^{n}(x_i-\bar{x})^2$ and that this also appears in our expression to calculate the standard deviation, or the variance, of x.

It will be recalled that we saw how we could rearrange the formula for the variance to make it easier to find using a calculator. So it is not really surprising to find that we can do the same thing for our expression for 'b'. The alternative formula may look more cumbersome at first sight, but is much more convenient for use, especially with a calculator, so we will always use it from now on. The formula is:

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

For calculating 'a' we retain the expression given above

$$a = \bar{y} - b\bar{x}$$

It will be seen that the formula for the straight line which best fits the data can be calculated if we work out values for Σx, Σy, Σx^2, Σxy and n.

We should make a cautionary note here that although this line is the 'best fit' for our data, according to the criteria we used, this does not mean that it is a perfect fit, or even that it is a good one. There will always be one straight line which fits the data better than any other straight line does, but whether this fit is a good one or a bad one, will depend on how scattered the series of paired observations were. Later in this topic we will develop a measure which will show us how well the line actually corresponds with the data.

Regression analysis is very extensively used in practice in estimating relationships between economic variables, such as demand and supply curves, relationships between income and expenditure, and so on. It also should prove very valuable in the analysis of fisheries statistics. Furthermore, it has great use in time series analysis, whereby we fit a straight trend line to data which is available, in order to estimate data which are missing, and most importantly in order to project forward to make estimates for future periods.

Let us look at one or two practical applications. We will be able to see how the regression line is actually calculated, and how it can be used to make estimates or forecasts.

First, let us revert to our data on boats used (which we denote x) and catch obtained (y) in the artisanal fishery (Table 5.1). We will calculate Σx, Σy, Σx^2 and Σxy, which we will need in order to calculate the coefficient 'b'. We already know that n = 10. We will also calculate Σy^2, which, although not used to calculate the regression coefficients, will be required for another calculation a little later.

TABLE 5.1 : CALCULATION OF THE REGRESSION COEFFICIENTS FOR DATA ON BOATS OPERATING AND CATCH OBTAINED

x	y	x^2	xy	y^2
12	590	144	7080	348100
15	820	225	12300	672400
10	330	100	3300	108900
12	740	144	8880	547600
18	900	324	16200	810000
14	660	196	9240	435600
6	240	36	1440	57600
15	650	225	9750	422500
16	850	256	13600	722500
9	470	81	4230	220900
127	6250	1731	86020	4346100

$\bar{x} = 127/10 = 12.7$ $\bar{y} = 6250/10 = 525$

Now we can substitute in our formula.

$$b = \frac{\Sigma xy - \frac{\Sigma x \, \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$= \frac{86020 - \frac{127 \times 6250}{10}}{1731 - \frac{(127)^2}{10}}$$

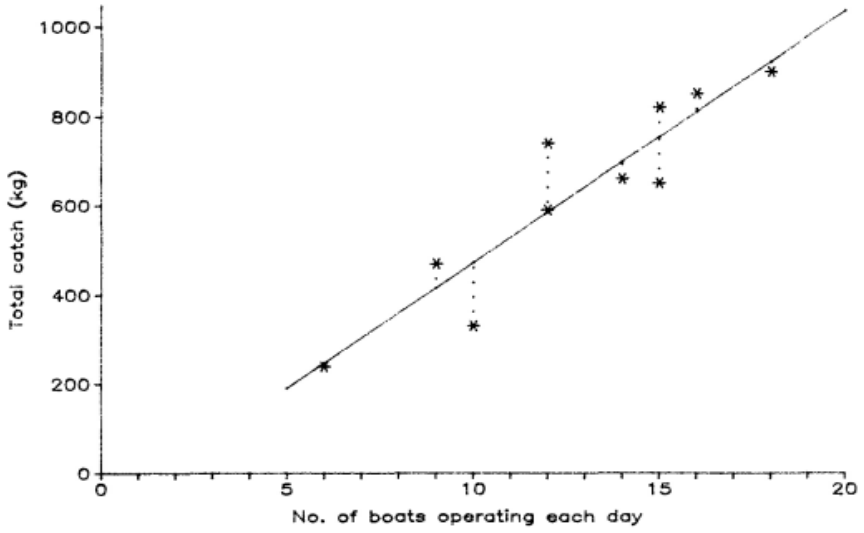
$$= \frac{86020 - 79375}{1731 - 1612.9} = 56.3$$

$$a = 625 - (56.3 \times 12.7) = -90$$

Therefore our regression formula is y=-90+56.3x.

In Figure 5.2, the regression line has been drawn in, and the vertical (y) deviation of each point on the scatter diagram from the regression line is also marked. What we have achieved in calculating the best fitting straight line to the data is to ensure mathematically that the sum of the squares of these 'y' deviations is the minimum possible. For any other line which we try to draw to fit the data, the sum of the squares of these deviations would be greater than for our regression line.

FIGURE 5.2 : CATCH PER DAY BY NUMBER OF BOATS OPERATING SHOWING LINE OF BEST FIT AND DEVIATIONS FROM THE LINE Y=-90 + 56.3 X



If now we want to estimate how much fish would be caught on a day when 14 boats are operating, we simply substitute 14 for x, and we have

$$y = -90 + (56.3 \times 14) = 698$$

So our equation estimates that 698 kg of fish would be caught. It is interesting to note from the actual data that there were 14 boats operating on one day, and the catch was 660 kg, which is very close to the estimate provided by our regression line.

We should also note that the regression line does not provide us with a good estimate for days when very few boats operate. For example, if there is only one boat operating, the equation predicts that the catch would be $-90+(56.3 \times 1) = -33.7$ kg, which is obviously nonsense. Our original data did not contain any observations for very small numbers of boats operating, so it is perhaps not so surprising that the equation is not good for making estimates when that situation arises.

Another way of interpreting the equation is to note that 'b', the coefficient of the slope of the line, equals 56.3. That means that the regression line estimates that the total daily catch will increase by 56.3 kg for every extra boat engaged. It tells us in effect: multiply number of boats by 56.3 kg, but then deduct 90 kg from the estimate this gives.

For our second illustration, we will look at a time series of annual data. We will use data on fish catch for Country ABC which we have plotted in Figure 2.2, but we will eliminate the observation for 1978. We assume that for some reason statistics were not recorded that year, and we want to use a regression equation to estimate what that year's catch would have been. We will also use the equation to forecast what catch can be expected in 1985 and 1986.

We note that 'year' is plotted on the x-axis in Figure 5.3. We could use 1976, 1977, etc. as values of x, but we would then have to deal with very large numbers. It is far easier to label 1976 as year 1, 1977 as year 2, and so on. This greatly simplifies calculations, and gives exactly the same answer for the predicted values of catch.

The catch figures were not actually listed out in Topic 2, but could be seen fairly accurately from the graph. The actual data, together with calculations of Σx^2, Σxy and Σy^2 (which we require for a later exercise) are shown in Table 5.2. We should note here that n=8, because with data for 1978 excluded from our calculations of the regression line, we have only 8 pairs of observations.

TABLE 5.2 : CALCULATIONS OF THE REGRESSION COEFFICIENTS FOR TOTAL ANNUAL FISH CATCH IN COUNTRY ABC

x	y	x^2	xy	y^2
1	604	1	604	364816
2	552	4	1104	304704
4	677	16	2708	458329
5	621	25	3105	385641
6	875	36	5250	765625
7	880	49	6160	774400
8	774	64	6192	599076
9	869	81	7821	755161
42	5852	276	32944	4407752
$\bar{x} = 42/8 = 5.25$			$\bar{y} = 5852/8 = 731.5$	

Substituting in our formula we have

$$b = \frac{32944 - \frac{42 \times 5852}{8}}{276 - \frac{1764}{8}}$$

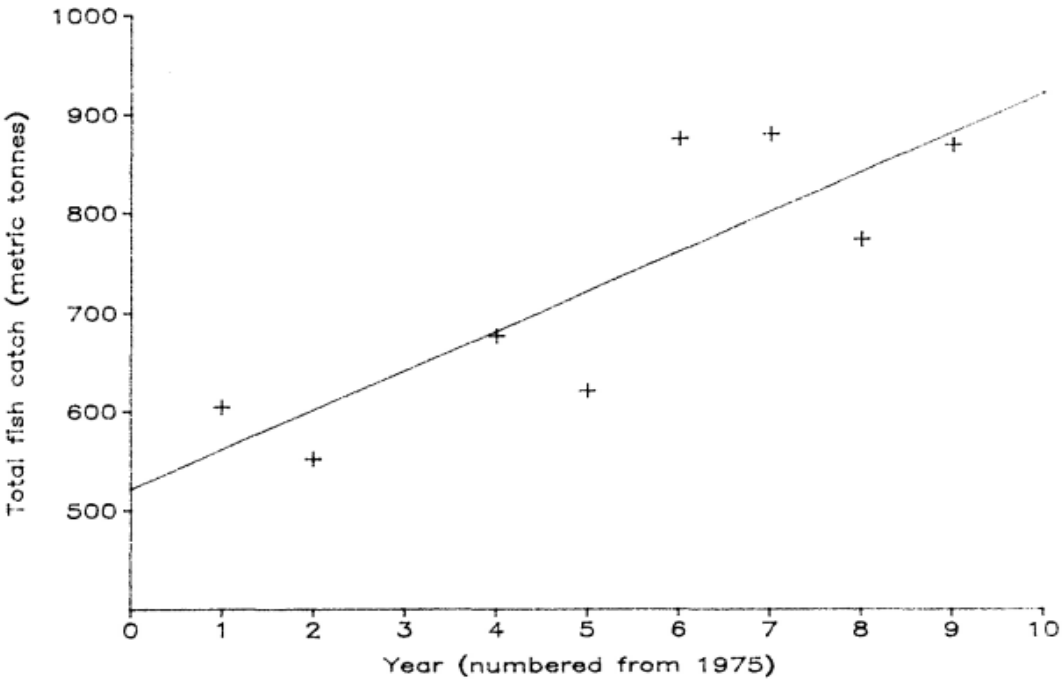
$$= \underline{40.0}$$

$$a = 731.5 - (40 \times 5.25)$$

$$= \underline{521.5}$$

Our equation therefore is y = 521.5+40x. This can be interpreted to mean that the trend line indicates estimated production of 521.5 tonnes in year 0 (i.e. 1975), increasing by 40 tonnes per year. Figure 5.3 shows this.

FIGURE 5.3 : TOTAL ANNUAL FISH CATCH IN COUNTRY ABC SHOWING LINE OF BEST FIT, Y=521.5 + 40.0 X



This allows us to make regression estimates for the other years, namely,

1978 (= year 3) y = 521.5 + 3 x 40 = 641.5 tonnes

1985 (= year 10) y = 521.5 + 10 x 40 = 921.5 tonnes
1986 (= year 11) y = 521.5 + 11 x 40 = 961.5 tonnes

In practice we would have to qualify these estimates by making allowance for factors other than the long-term trend. No doubt weather conditions and other factors are going to have a substantial influence on the actual catch obtained. We can readily see from the graph that 1980 was quite a bad year, and 1976 and 1981 were relatively good years. So we should say that "without making any allowance for outside influences" our regression line makes estimates of annual catch as we showed above. We may hope that local knowledge, or other recorded information, would enable the regression estimates to be adjusted to take account of these outside influences.

We can see indeed from Figure 2.2, which included a value for 1978, that that year was a relatively poor one: catch was well below that recorded in 1976, for instance, and our regression equation is predicting an increase of 40 tonnes every year. If no allowance is made for these external factors (such as weather) and the estimate for 1978 catch is made solely on the basis of the regression equation, then the level of the catch will be over-estimated.

Linear regression estimates are not magical numbers which show exactly what would occur in given situations. They are simply best estimates based on certain available data, and without taking account of any data other than those relating to the two variables used in the regression analysis. Obviously the closer each paired observation lies to the regression line, the more accurate the estimate is likely to be, and the greater the confidence we can have in our estimates. A little later we will develop simple estimates which will indicate how closely our regression line fits the available data.

One final point should be made. Because our regression line in a time series is only an estimate, and because it takes into account only movement in one variable against time, it is of limited value for making forecasts well into the future. In our illustration we projected forward for two years, and perhaps that is as far as we should go. The further we attempt to project beyond the points in our data set, the less reliability we can place on those projections.

5.3 Non-linear relationships

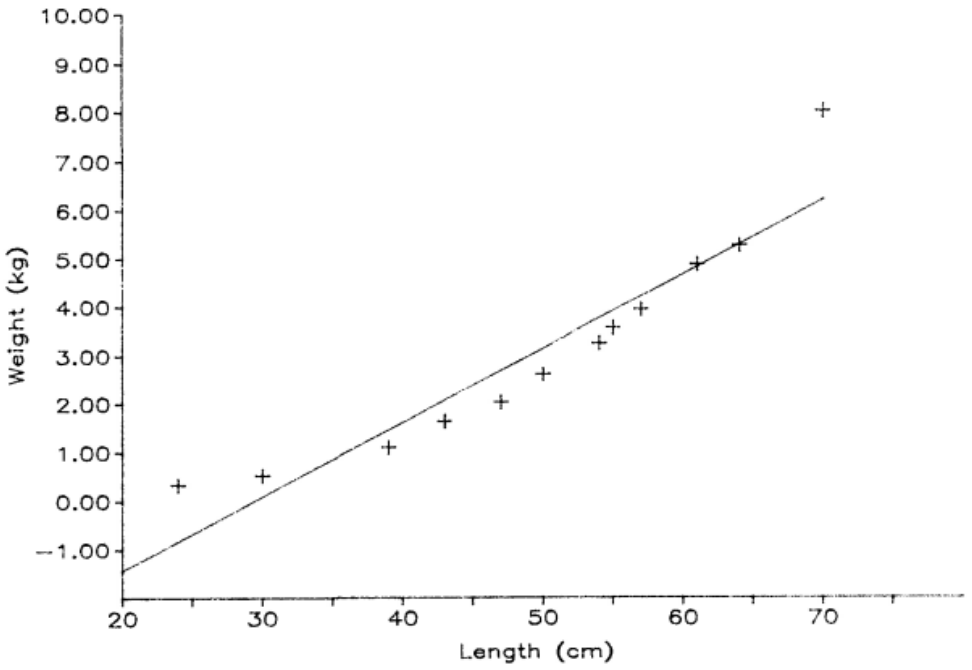
So far we have examined situations where we could reasonably expect to fit a straight line to the available data. All the points lie reasonably close to the regression line we were able to establish.

However, often two variables will bear a clearly non-linear relationship to each other. For instance, it may seem apparent from a visual inspection of a scatter diagram that the points seem to lie more or less along a curve. If we look at the length-weight relationship of skipjack, for instance, it is apparent that the observations (with length plotted on the x-axis and weight on the y-axis) clearly follow a curved path, which slopes upwards more steeply at the right hand side of the graph. In other words, for larger fish there is a large increase in weight for a relatively small increase in length.

Figure 5.4 is a scatter diagram of the length and weight of a sample of 12 skipjack, with points showing the curved pattern we would expect. This does not mean we cannot find a linear regression function which best fits these points. We can do so; there is always one line which fits a set of paired observations better than any other line will. In fact, a quick calculation will show that the equation of that line is $y = -4.5+0.153x$,

and this is plotted on the diagram. It is apparent that the regression line lies above the observed points for skipjack between about 35 and 60 cm in length, and is well below the observation for very small and very large skipjack.

FIGURE 5.4 : WEIGHT BY LENGTH OF SKIPJACK SHOWING LINE OF BEST LINEAR FIT, Y=-4.5 + .153 X



If we substitute in the equation for a skipjack of length 28 cm, we find a predicted value of y (weight) of -0.22 kg, which is clearly nonsense.

In such a situation we can sense that there must be a better way to approach the problem. In our example the points lie so close to a curve we can sketch in that we should be able to find an equation which gives a very good fit for our data.

In fact, it has been established that a relationship between weight (y) and length (x) of fish conforms to the equation $y = a.x^b$. Taking the logarithm of this equation, we have

$$\log_e y = \log_e a + b \log_e x$$

and this is a relationship similar to our equation for a straight line $y = a+bx$. In other words, if we plot a scatter diagram of log y against log x, we should be able to derive values of the coefficients log a (the intercept of the y-axis) and b (the slope), so that we have a regression line which will provide a better fit for our data than the line $y = -4.5+0.153x$ which we drew previously.

The calculations are shown in Table 5.3 and the graph with the appropriate regression line plotted is in Figure 5.5.

FIGURE 5.5 : LN (weight) BY LN (length) OF SKIPJACK SHOWING LINE OF BEST FIT, Ln (Y) = -10.82 + 3.01 Ln (X)

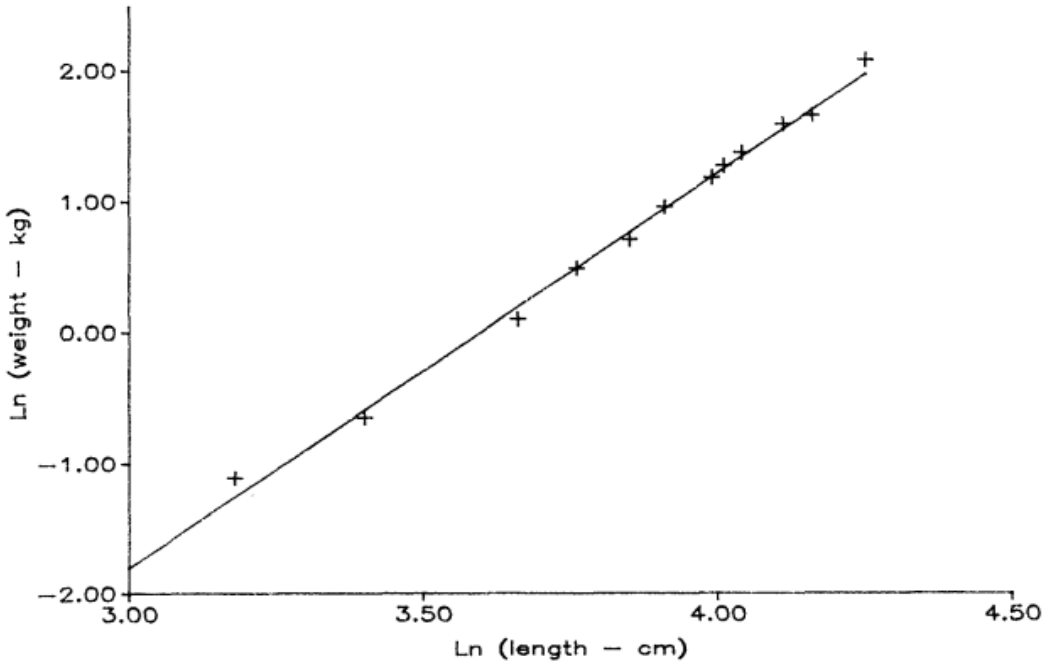


TABLE 5.3 : LENGTH AND WEIGHT DATA FOR A SAMPLE OF SKIPJACK: LOGARITHMIC RELATIONSHIP

Length (cm) x	Log x	Weight (kg) y	Log y
24	3.18	0.33	-1.11
30	3.40	0.52	-0.65
39	3.66	1.10	0.10
43	3.76	1.64	0.49
47	3.85	2.03	0.71
50	3.91	2.61	0.96
54	3.99	3.25	1.18
55	4.01	3.56	1.27
57	4.04	3.94	1.37
61	4.11	4.88	1.59
64	4.16	5.27	1.66
70	4.25	8.01	2.08

$\sum(\log x) = 46.32$ $\sum(\log y) = 9.65$

$\sum(\log x)^2 = 179.9$

$\sum(\log x)(\log y) = 40.57$

and by substituting in our formula for the regression coefficients we find

$$\log y = -10.82 + 3.01 \log x$$

(or $\log y = \log 0.00002 + 3.01 \log x$)

Now if we try this relationship for various values of x, we find the following:

If length = 28 cm, $\log x = 3.33$

$$\log y = -10.82 + (2.01)(3.33)$$
$$= -0.797$$
$$\therefore y = 0.45 \text{ kg}$$

This is obviously a much better estimate than the previous equation provided.

Similarly, if length = 43 cm, $\log x = 3.76$

$$\log y = -10.82 + (3.01)(3.76)$$
$$= 0.498$$
$$\therefore y = 1.65 \text{ kg}$$

and we may observe that this is almost identical with the weight (1.64 kg) of the 43 cm skipjack in our original data set.

In the next section we will provide a measure which clearly shows that this regression line is a far better fit to our data than the first simple formula we derived.

As a generalisation, we can say that when a series of paired observations appearS to follow a simple curve, we should be able to establish some relationship which will permit a much better straight regression line to be drawn than will be obtained by the basic formula of a line, $y = a+bx$. This may involve logarithmic, square, square root or other functions of x and/or y. There are exceptions to this: the relationship between length (or weight) and age of fish is an example where it has been found that the relationship cannot be ´linearised´. In that case, analysis has to be undertaken using techniques which are far more complex than we have discussed here. However, the idea of finding a linear relationship from basic data which is non-linear is a very important one in the analysis of fisheries statistics.

5.4 How well does the mathematical relationship describe the data

In the previous section we have seen how we can fit a straight line to a set of data, where we have n measurements of some independent variable x, and n associated measurements of some dependent variable y. We used a straight line because it is the simplest mathematical relationship we can find, and we have seen that we can still sometimes fit a straight regression line to data which bear a non-linear relationship. It is obvious, however, that for some data sets a straight line is not a good way to describe the data. If we look, for example, at Table 5.4 and Figure 5.6 we can see here that there seems to be very little relationship between the variables.

TABLE 5.4 : EXAMPLE OF NO CLEAR RELATIONSHIP BETWEEN VARIABLES. Catch per unit effort (i.e. catch per boat days fished) by effort (boat days fished) - series of 12 monthly observations.

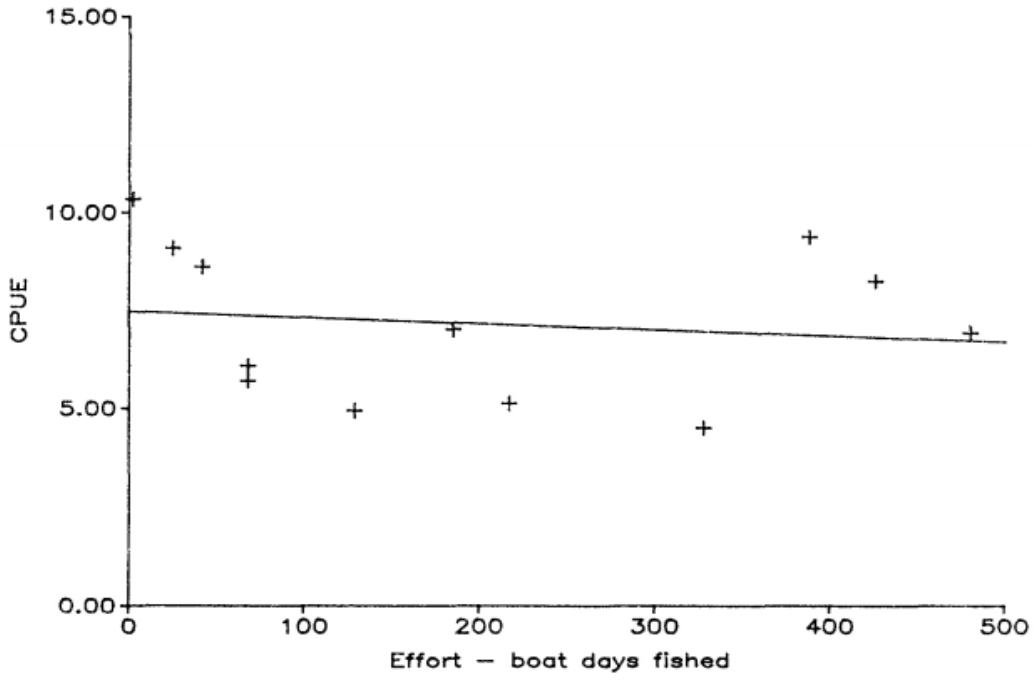
Effort (x)	CPUE (y)	x^2	xy	y^2
129	4.94	16641	637.26	24.40
328	4.52	107584	1482.56	20.43
217	5.14	47089	1115.38	26.42
68	6.12	4624	416.16	37.45
25	9.11	625	227.75	82.99
2	10.35	4	20.70	107.12
68	5.72	4624	388.96	32.72
42	8.64	1764	362.88	74.65
388	9.40	150544	3647.20	88.36
426	8.27	181476	3523.02	68.39
185	7.05	34225	1304.25	49.70
480	6.95	230400	3336.00	48.30
2358	86.21	779600	16462.12	660.95

Substituting in our formula as usual, we find a = 7.48 and b = -0.0015. The regression line therefore becomes

$$y = 7.48 - 0.0015x$$

This is drawn on the scatter diagram, Figure 5.6.

FIGURE 5.6 : CPUE (CATCH PER BOAT DAY FISHED) BY EFFORT SHOWING LINE OF BEST LINEAR FIT, Y = 7.48 - 0.0015 X



We may observe that the coefficient of the slope is negative, and therefore the line slopes downwards to the right. This indicates that the higher the effort in terms of boat-days fished, the lower the return in terms of catch per boat day.

We can see, therefore, that we can always find the equation of a line for almost any set of data. All we have to do is to perform the various calculations and put the values in the equations for ´a´ and ´b´. But this is not really sufficient; we have to guarantee that when we have calculated the line for the data that it will provide reliable predictions. Because of the way we have calculated the values of a and b we know that this will be the best line for these data, but we do not know if the data points are closely grouped around the line, or if they are widely dispersed. In Figures 5.2 and 5.3 the data points do seem to be fairly closely distributed around the line we have calculated, but in Figure 5.6 we can see at a glance that this is not so; only 2 of the 12 points lie anywhere near the regression line we have drawn. Common sense would tell us that the predictions in the first two cases are much more likely to be reliable than in the third. What we need therefore is not only a method for finding the best equation of a line through the data, but also some measure of how close the data points are to the line.

5.5 The coefficients of correlation

If we look at any of the scatter diagrams above, we can see that the points do not lie exactly on the line we have drawn. Since we are interested in predicting the y values we can see how far each point is from the regression line in the y direction, that is, vertically. We call the vertical distance from the line to any of the data points the ´residual´. In effect, we can say that each observed value y_i is equal to a value ´$a+bx_i$´ plus the residual. The smaller the residuals, the closer the points are to the line, and the better the line ´fits´ the data. One way, therefore, to see how close our data points are to the line is to measure the residuals. We can do this graphically, but we can also do it mathematically by calculating a value known as the correlation between x and y. The correlation is a measure of how close the relationship between x and y is to a straight line.

Before we go on to see how we can calculate the correlation, it is necessary to go back and look at the expression,

$$\sum_{i=1}^{n} (x_i - \bar{x})\ (y_i - \bar{y})$$

which we used when calculating the slope, b, of the regression line. From Topic 4 we remember that to measure the variation among a set of data points x_i, x_2 ...x_n we can calculate the standard deviation or the variance. The formula for the variance was given by:

$$\sum_{i=1}^{n} (x_i - \bar{x})\ /\ (n - 1)$$

Now this looks somewhat similar to the expression we have just written down. This can be looked at as measuring the joint variation of x and y about their respective means. We call this quantity,

$$\sum_{i=1}^{n} (x_i - \bar{x})\ (y_i - \bar{y})\ /\ (n - 1)\ ,\quad \text{or}\quad \sum (x - \bar{x})\ (y - \bar{y})\ /\ (n - 1)$$

the co-variance of x and y. It shows how the two variables change together. If they are closely related, this value will be high; if they are not closely related, it will be small. Notice, however, that the

co-variance can be negative; if the relationship slopes downwards then it will be less than zero.

The correlation between x and y is measured by a coefficient, which we denote as 'r', and this is given by the co-variance of x and y divided by the product of the standard deviations of x and y. In terms of a mathematical formula we can write it as:

$$r = \frac{\Sigma(x - \bar{x})\ (y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \ .\ \Sigma(y - \bar{y})^2}}$$

(The values (n-1) can be divided out in both numerator and denominator). If we only used the co-variance to measure the relationship we would have problems in comparing different sets of data. For example, if we changed the units of measurement of one, or both, of the variables, we would change the value of the co-variance. We get round this problem by dividing by the product of the standard deviations; this means that r can only be a value between plus and minus one. If all the data points lie exactly on an upward sloping line, then r will be +1; if they all lie on a downward sloping line, r will be -1. Values in between, then, tell us how strong the relationship is between x and y.

If r is very close to +1, we say there is a strong positive correlation: y increases as x increases, and the relationship is good. If r is close to -1, there is a strong negative correlation: y decreases as x increases. When r is close to zero (either positive or negative) there is very little relationship between the two variables.

As with all similar measures we have studied, we find that in practice our alternative version of the formula is more suitable for ordinary use, especially with a calculator. This is

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \ .\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

We may note that r can be calculated if we can obtain the values of n, Σx, Σy, Σx^2, Σxy, all of which were used in our earlier calculations of the regression coefficient 'b', and one additional value, Σy^2. It was for this purpose that we obtained the value of Σy^2 in our earlier examples in this topic.

If we revert to our data in Table 5.1, and substitute in our formula for 'r', we have

$$r = \frac{10(86020) - 127(6250)}{\sqrt{10(1731) - (127)^2} \ .\ \sqrt{10(4346100) - (6250)^2}}$$

$$= \frac{66450}{\sqrt{1181} \ .\ \sqrt{4398500}} = 0.92$$

This is a positive value, quite close to +1, so we can say that there is a good positive relationship between the two variables, according to our data.

We can go through a similar process to calculate 'r' for our time series data on annual fish catch (Table 5.2), and we find r = 0.84. In other words there is a good positive correlation (i.e. our catch is moving

upwards over time), but the relationship is not as strong as in the previous example.

The calculation of the coefficient of correlation for our two alternative formulae for the length-weight relationship of skipjack is quite interesting. We would expect a strong positive relationship, because obviously weight increases as length increases. In fact, in our first equation, which was $y = -4.5+0.153x$, we can calculate that $r = 0.94$. This shows that, despite the poor estimates which the regression line gave for very small and very large fish, the fit of this line to the data we had available is good.

However, when we examine the second equation, $\log y = \log (0.00002)+3.01 \log x$, we find that $r = 0.997$. This is very close to 1, and clearly shows that we were able to find a much better regression line by our special technique of using a 'log log' relationship.

Finally, we can look at the relationship of 'effort' to 'catch per unit effort', portrayed in Table 5.4 and Figure 5.6. We noted at the time that y was decreasing as x increased, so we must expect r to be negative; we also observed that the relationship appeared to be very weak, so we should anticipate obtaining a value for r which is closer to 0 than to -1. When we make the appropriate calculation we find that $r=-0.13$, which is so close to zero that hardly any relationship at all can be established. We could have very little confidence at all in any conclusions we attempted to draw from this regression equation.

We can summarise what we have been discussing by saying that the regression coefficients measure the linear relationship between two variables, and the correlation coefficient tells us how closely the data fit this relationship. The two are clearly related to each other, but it is important not to confuse them because they measure different things.

It is possible to go much further than this in analysis, and calculate 'confidence limits' for our estimates. Essentially this is another way of expressing the goodness of fit of a relationship, but this is beyond the scope of this course. However, we may note that in our next topic, Sampling, we will be calculating confidence limits for estimates derived from samples, and there is a close parallel between the two.

5.6 Seasonal variation

Earlier in this topic, in examining links between two variables, we looked at an illustration of one of the most important relationships - that between the values of a characteristic and time. This relationship is referred to as a time series .

That particular example concerned a time series of annual data, and we obtained a regression line which best fitted the observations and we referred to this as the trend line. The line endeavours to show how the value of the characteristic is changing in the long term.

However, often we are vitally interested, not just in the trend in annual values, but in movements in the shorter term - from month to month, for example. We then may find that our study is complicated by a pattern of peaks and troughs in the value of our observations. If this pattern tends to repeat itself each year, we call this seasonal variation. By this we mean that the value of our variable tends to vary according to the time of the year; at some times it will be high, at others low, but the pattern

will tend to repeat itself regularly. This kind of variation occurs quite often in time series, particularly those related to production or climatic factors.

If we look at the figures of tuna catch in Fiji given in Table 5.5, we will see that there is a very high, and quite regular, seasonal variation. In each of the four years covered by our data, the catch is highest in the first five months of the year, from January to May, and is lowest around August to October, being almost zero each September.

TABLE 5.5 : IKA CORPORATION, FIJI - ESTIMATED TUNA CATCH, MONTHLY 1979-1982. Calculation of 12-month moving averages.

Month	Tuna Catch (tonnes)	12-month Total	24-month Total	Moving Average
Jan. 79	594			
	488			
	535			
	468			
	566			
	354			
	190	3358		
	18	3094	6456	269
	2	2980	6074	253
	0	2911	5891	245
	57	2751	5662	236
	86	2257	5008	209
Jan. 80	330	1988	4245	177
	374	1927	3915	163
	466	2009	3936	164
	308	2038	4047	169
	72	2123	4161	173
	85	2205	4328	180
	129	2335	4540	189
	100	2824	5159	215
	31	3408	6232	260
	85	4001	7409	309
	139	4365	8366	349
	216	4888	9253	386
Jan 81	819	5267	10155	423
	958	5568	10835	451
	1059	5553	11121	463
	672	5522	11075	461
	595	5522	11044	460
	464	5542	11064	461
	430	5657	11199	467
	85	5467	11124	464
	-	5288	10755	448
	85	4991	10279	428
	159	4892	9883	412
	331	5015	9907	413
Jan. 82	629	4939	9954	415
	779	4659	9598	400
	762	4624	9283	387
	573	4624	9248	385
	718	4666	9290	387
	388	4618	9284	387
	150	4555	9173	382
	50			
	-			
	127			
	111			
	268			

Source: Annual Report 1982, Fisheries Division, Ministry of Agriculture and Fisheries, Fiji

We would like to look beyond this seasonal variation, and try to find how the catch, as a whole, is changing over time. Obviously we cannot just compare data for consecutive months. It would be quite unreasonable to conclude, for example, that the tuna catch is falling, just because the amount caught fell each month from May to September 1982. The catch falls in that period every year, and what we would need to establish, in order to get any picture of a longer-term trend, is whether the fall in these months of 1982 was greater or less than the fall which is usually recorded at that time of the year.

There are various techniques available to calculate this seasonal variation. If we can obtain a measure of this variation we can eliminate it from our data, to give us a more meaningful trend line. This process is referred to as seasonal adjustment. There are now quite sophisticated computer programs which are widely used by statisticians all over the world, to seasonally adjust (or 'deseasonalise' as it is sometimes called) any time series data. We will not go into this topic in detail during this course, but we will look at the first step in the process, the moving average, and see how this can assist in eliminating seasonal patterns from data, in order to highlight the trend.

To see how this works consider the following sets of numbers:

4, 5, 7, 3, 6, 4, 5, 3, 7, 6, 3, 4.

The average (arithmetic mean) of the first three numbers is: (4+5+7)/3 = 5.3. We could then 'move' the average along, and find the average of the second three numbers, 5, 7 and 3; this will be (5+7+3)/3 = 5.0. We can repeat the process by moving the average along the series, one observation at a time. This then gives us the following situation.

Original series	4	5	7	3	6	4	5	3	7	6	3	4
Moving average of order 3		5.3	5.0	5.3	4.3	5.0	4.0	5.0	5.3	5.3	4.3	

Since our first moving average is the mean of the first three terms (4, 5 and 7) we can place this underneath the middle value and then move the average along one value each time. We have calculated a moving average of three terms; we call this an average of order 3. An average of order 5 would include 5 terms and so on.

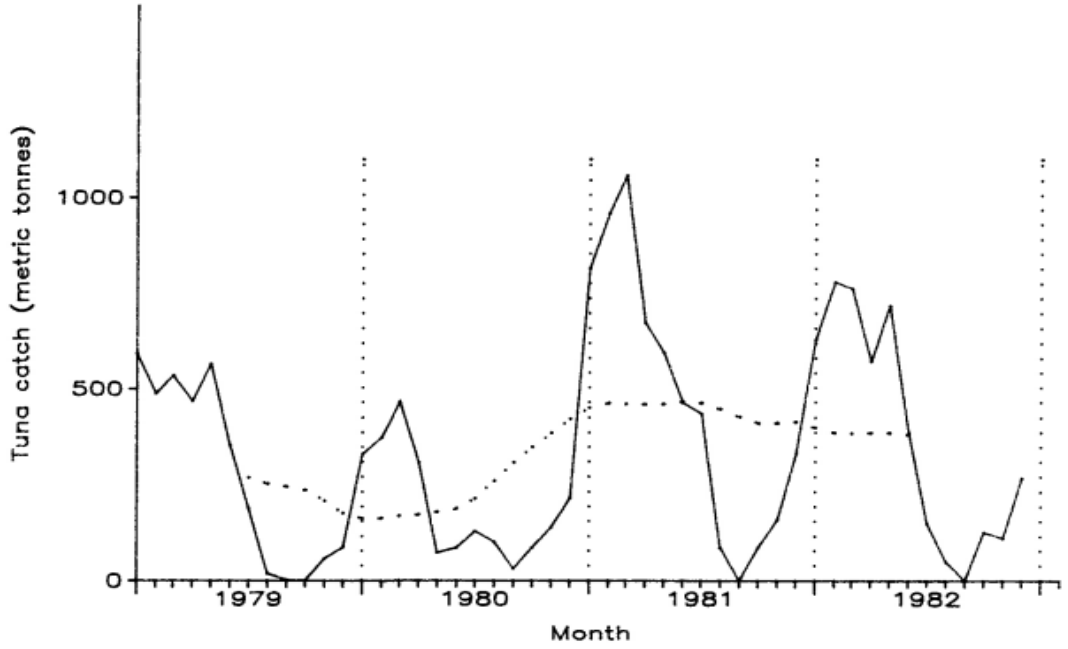
One of the reasons that we calculate a moving average is to reduce the variation in the original series; in our example the moving average is less variable than the original series. We also see that the series of averages is shorter than the original; we have "lost" terms at the beginning and the end. This is inevitable, because of the way we calculate the moving average. The greater the order, the more variation will be smoothed out, but the more terms that will be "lost" at the beginning and end of the series.

A moving average, therefore, will smooth out random variation in a time series, and if we choose an appropriate order it will also eliminate the seasonal variation. What will be left will be the trend values. We have defined the seasonal variation to be that which varies with the seasons, so which will repeat itself annually. If we have monthly data, therefore, we shall need a moving average of order 12 to eliminate the

seasonal variation; with quarterly data we will use an average of order 4. There is, however, one extra problem when we use an average with an order which is an even number. If we calculate an average of order 12 and start with observations from January 1979 to December 1979 (as is done in Table 5.5), this will be centred half-way between June and July 1979. The average for February 1979 to January 1980 will be half-way between July and August 1979, and so on. This is obviously inconvenient; we want to know the trend value of June and July and not some mid-point. What we have to do, once we have calculated the first average, is take a second average of two terms which will 'centre' our trend values. We call the resulting moving average as one of order 2x12.

The calculations of the 2x12 moving average for the Fiji tuna catch were shown in Table 5.5; the moving averages, or trend values, have been plotted on Figure 5.7. We can see that the moving average has removed most of the random and seasonal variation and so allows us to get a much better idea of the trend.

FIGURE 5.7 : MONTHLY TUNA CATCH (FIJI) AND 12-MONTH MOVING AVERAGE (broken line)



Of course moving averages do not eliminate the effects of different conditions from year to year: in fact they help to highlight these effects. It is very easy to see from the moving average line on the graph that 1980 was a bad year, and 1981 a good year, for example.

The example we have given relates to repeated fluctuations which occur during a year, and this is probably the most common seasonal variation we will encounter. However, we can also encounter seasonal patterns over different periods - for example:

Monthly

There may be a repeating pattern during each lunar month. Studies have shown that catches of baitfish are regularly highest at the time of new moon, and lowest at time of full moon.

Daily

Some variables may change regularly at different times of the day. The price of fish in the local market, for instance, may be highest early in the morning, and may be lower later in the day as vendors reduce their prices in order to get rid of their unsold stock.

The same techniques can be applied in these circumstances, in order to remove the repeating pattern, and to obtain a more realistic trend line for data.