Assumptions

Regression relies on more assumptions than any other technique discussed in this book. Yet, frequently regression is applied in geography and elsewhere with an apparent disregard for, or even ignorance of, the assumptions. For this reason they will be discussed in some detail here. If the assumptions are not fairly well satisfied, inferences made from a regression may be invalid, although the regression equation may still be of value in describing the relationship between two variables.

There are five major assumptions:

(1) The independent variable x is not a set of sample values. Strictly speaking the values of x should be fixed values chosen by the researcher, and not sample measurements. If this is not the case, as in the example used here, then the values of the independent variable must at least have been measured with a negligible amount of error.

(2) The values of the dependent variable are normally distributed.

(3) The variance of the dependent variable is constant for all values of the independent variable.

(4) The value of the *residuals* (the differences between the observed and predicted values of the dependent variable) have a normal distribution.

(5) The values of the residuals are independent of each other, i.e. they are randomly arranged along the regression line.

Since these assumptions are often ignored in geographical applications of regression, it is worth emphasizing them by considering a hypothetical situation in which they are all satisfied.

In a particular lake, samples of water are taken at depth intervals of one metre, and the salinity of each sample is measured. Figure 6.7a is a graph of the data obtained, with the least squares regression line drawn through them. The independent variable in this case is depth, which is not a sample variable, but one which has been fixed at certain specified values. It is not subject to sampling error (since a metre is always a metre) and can be measured with a high degree of accuracy. Assumption 1 is therefore satisfied.

The dependent variable, salinity, is a series of sample measurements, which are subject to sample variation. Taking a sample of water at each depth implies selecting a particular 'lump' of water. Since the water is constantly in motion the selection of any particular 'lump' is a chance happening. Assumption 2 means that if the researcher were to take, say, 1000 samples at any one depth, and produce a frequency distribution of the salinity values from these samples, it would be a normal distribution.

Assumption 3 means that if this same process were repeated at each depth, the shapes of all the frequency distributions would be identical. There would however, be a difference between the means of these frequency distributions. In fact the least-squares technique assumes that the means of all the distributions lie exactly along the regression line. Figure 6.7a illustrates assumptions 2 and 3 diagramatically. Figure 6.7b shows the residuals, the deviations of the observed from the expected values of the dependent variable. For assumption 4 to be satisfied, the frequency distribution of the residuals should be normal. According to assumption 5, there should be no pattern in the residuals. They should have all the characteristics of a sequence of random numbers.

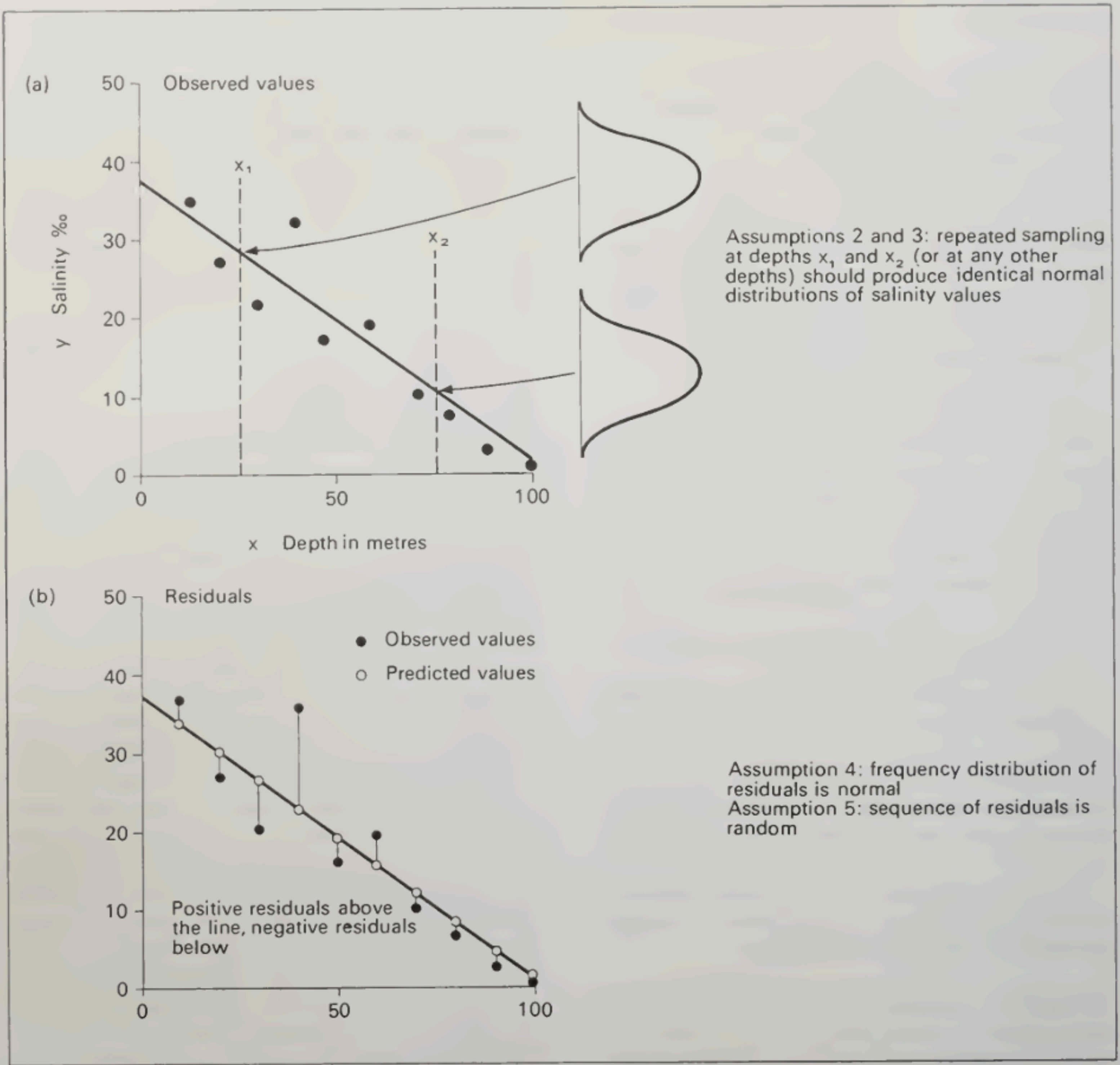
So far, regression has been discussed as a method for calculating an equation which can be used as a concise description of form of a relationship between two variables. If the data relate to a sample study, as they frequently do, the researcher needs to know two things:

(1) How well the regression line fits the data.

(2) How accurate any predictions based on the regression, are likely to be.

Goodness of Fit

The equation linking Fahrenheit and centigrade temperatures (Equation 6.1) is clearly a complete explanation of the relationship between the two temperature scales. Equation 6.5, on the other hand, does not provide a complete explanation of the relationship between moisture content and number of alders. The regression line, shown in Figure 6.6, does not pass through any of the observed points. In other words, the predicted value of the dependent variable ($\hat{y}$) is different from the



Assumptions 2 and 3: repeated sampling at depths x_1 and x_2 (or at any other depths) should produce identical normal distributions of salinity values

Positive residuals above the line, negative residuals below

Assumption 4: frequency distribution of residuals is normal
Assumption 5: sequence of residuals is random

Figure 6.7 Assumptions of regression