PREDICTION, LEAST SQUARES ESTIMATION AND REGRESSION

Although correlation is related to prediction, as has already been pointed out, the correlation coefficient in itself does not permit statements about the relationship between the scores of specific individuals on the X and Y variables, but only about common variance. In other words, general statements can be made about tendencies in a group of subjects, but not about individuals. We turn now to the problem of predicting individual scores.

When specific predictions of the scores of individuals are to be made, regression, is the appropriate procedure. The typical regression problem has the following form: We know the scores of individual members of a group of subjects on variable X, and wish to predict each person's score on variable Y. We also know the scores of some subjects on both X and Y (but obviously not those of the people for whom we wish to predict Y). Using calculations based on the group for whom we know both X and Y, it is possible to calculate a regression coefficient. This can then be applied to the X scores of the other subjects to predict their Y scores. This procedure would be useful in the following example:

Estimation

As has already been pointed out, perfect prediction of scores on Y from scores on X is only possible when X and Y correlate perfectly: usually when r = 1.00, but theoretically also when r = -1.00. Otherwise, we can only estimate the Y scores. In other words, the regression coefficient only permits estimates of the Y scores.

Least squares solutions

The approach adopted in statistics for, calculating the best estimate of b is the method of least squares. Least squares solutions are by no means confined to problems in the area of regression; you have already encountered a least squares approach earlier, for instance in calculating the arithmetic mean or the standard deviation: The arithmetic mean of a distribution is that point in the distribution (X), about which the sum of the squared deviations of the individual scores is a minimum; i.e., When X is the arithmetic mean,

$$E(x_i - \bar{x})^2$$

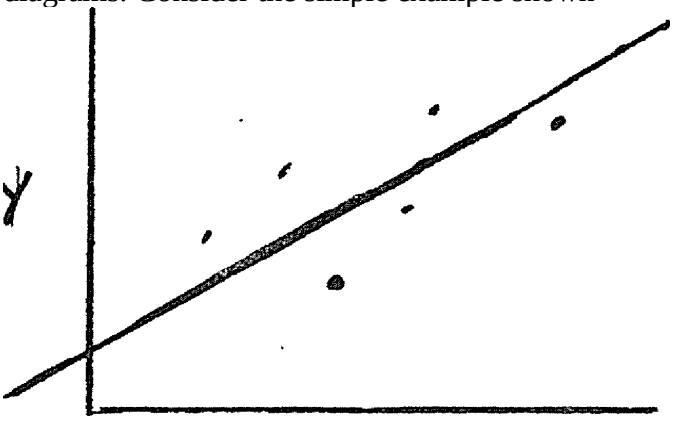
is at a minimum. (When the value of this summation is divided by the number of individuals in the distribution (N), the result is the variance, i.e., variance is itself a kind of least squares solution.)

In prediction problems, the least squares approach is applied to obtaining the best estimate of b in the following way: We take the group for which we already know both X and Y and -- in theory -- try out all possible values of b. In practice there are mathematical shortcuts, but in order to illustrate

the point here it is useful to pretend that all possible values of b are tried out. Every value of b which is tried out produces a set of predicted values of Y.

Regression

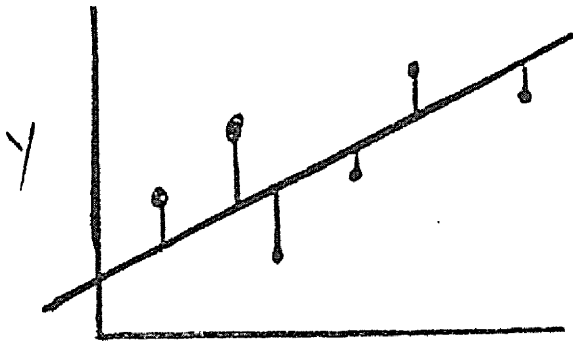
The easiest way of understanding regression is to think back to scatter diagrams. Consider the simple example shown



The cloud of points defines an ellipse (as is always the case with correlated variables).

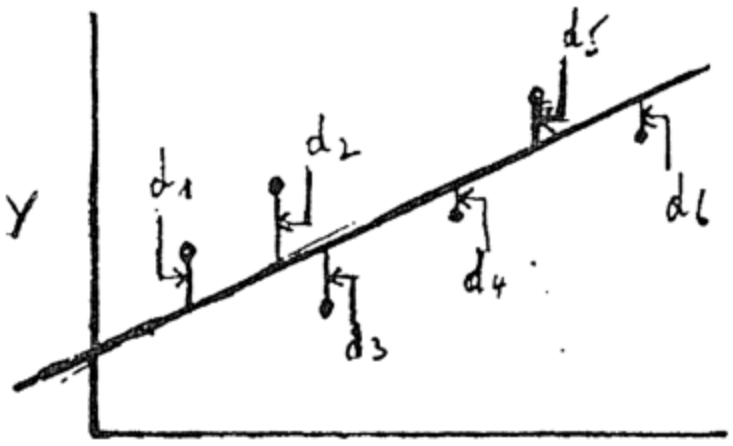
The problem in regression is to place a straight line through the cloud which best represents the cloud. This line is the regression line. The best line must be placed in such a way that it represents as closely as possible the line on which all the points would lie if they described a perfectly straight line. When any line is placed through the cloud, whether it is the "best" one or not, the distance of the points from the regression line can be represented by lines joining the points and the regression line, parallel to the y-axis. (The distance can also be represented by lines joining the individual points to the regression line, parallel to the x-axis or at right angles to the regression line. However, these lines are irrelevant for the present discussion.) Figure 9.2 shows the lines joining the points to the regression line, parallel to the y-axis.

Figure 9.2: Distances of individual points from the "regression" line.



The "best" position of the "regression" line is the one which you might have deduced - the one where the sum of the squares of the distances of the points from the regression line is at a minimum. This is the line of best fit or, strictly speaking the regression line. **I** referred above to any line representing the cloud of points as the "regression" line - this was incorrect, but simplified the discussion. Referring back to Figure 9.2, the line of best fit (the regression line) is the one where the sum of the distances marked d_1, d_2, d_3, d_4, d_5 and d_6 squared is at a minimum (see Figure 9.3).

Figure 9.3: Distances from the regression line



$E(d)^2$ is at a minimum

The regression coefficient is the slope of the line of best fit. For those who have done no coordinate geometry, the slope can be calculated by taking any pair of points in the cloud and calculating the following:

$$b = Y \; Y_1$$

At this point, one further important element in _regression can be introduced: if you look at the regression line depicted in Figures 9.1, 9.2 and 9.3 you will see that it cuts the y-axis. The value of Y at which the regression line cuts the y-axis is referred to as the Y-intercept, and has the abbreviation a. The approximate regression equation mentioned on p.50, can now be stated in its exact form; not $Y_i = bX_i$ as I have used up to now for ease of explanation, but:

$$Y_i = a + bX,$$

What this means is that to make the best estimate of Y-scores from X scores, it is necessary to know (1) the regression coefficient (b) and (2) the Y-intercept of the regression line (a).

To take a concrete example, return to the data in Table 8.1 on p.34, showing VCE scores and IQs. Suppose that we want to predict VCEs from IQ, i.e., IQ is the X variable and VCE is the Y variable. Suppose that the regression coefficient had been calculated and found to be 1.3 and the Y-intercept 150. The equation for estimating VCE from IQ with the highest degree of accuracy would then be:

estimate of VCE = 150 + 1.3 times IQ

Person 1 had an IQ of 117. The estimate of VCE would thus be 150 + 1.3 x 117. This equals 302. The actual value was 298, so that there was an error of estimate of 4. In the case of person 2, the estimate of VCE would be 150 + 1.3 x 132, which equals 322, i.e., an error of 26. If this calculation were made for all subjects in Table 8.1, the errors squared, and the squared errors added up, the resulting sum would be the smallest possible for any of the infinitely large number of combinations of possible a and b values (least squares solution). If this sum of squares were 0.00, the prediction of VCE from IQ would be perfect, it were very large (despite being the smallest possible), the prediction of VCE (or any other Y variable) from IQ (or any other X variable) would be poor. Intermediate sizes of the sum of residuals squared indicate intermediate degrees of usefulness of X for predicting Y.