than correlation techniques which are used less frequently than they once were. We shall consider regression first.

Linear Regression and the Method of Least Squares

As a typical regression problem consider the data in Table 22 on the comb-growth (increase in length+height of the comb) in 5 groups of 5 capons (castrated cocks) receiving different doses of androsterone (male sex hormone) (Greenwood *et al.*, 1935). It will be seen from Fig. 30, in which comb-growth is plotted against the logarithm of the dose, that there is an approximately linear relationship between these two quantities over the range of doses used. Comb-growth is obviously the dependent, and dose of androsterone the independent, variable.

Table 22

Comb-growth in capons receiving different doses of androsterone

Dose (mg. androsterone)	$\frac{1}{2}$	1	2	4	8
Log_2 dose (x)	-1	0	1	2	3
Comb-growth (mm.) (y)	8	5	13	17	17
	1	6	7	14	17
	1	9	12	14	20
	3	7	10	19	18
	1	4	11	13	15

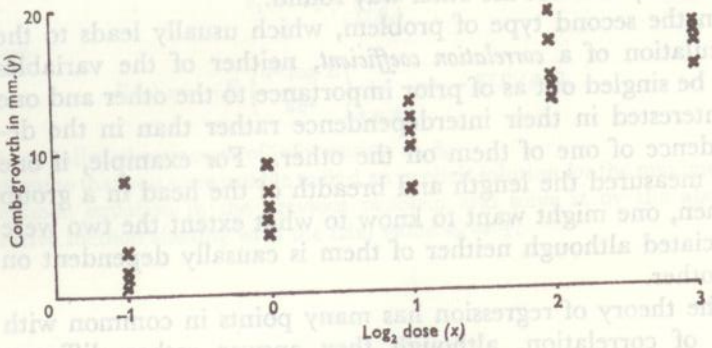

Fig. 30. Comb-growth in capons receiving different doses of androsterone

It is clear that, for a fixed value of the dose, the comb-growth varies considerably from one bird to another and may be regarded as a random variable with a mean, a variance and so on, which will be symbolised by $E(Y \mid x)$, $V(Y \mid x)$, etc., where Y stands for the comb-growth (the dependent variable) and x for $\log_2$ dose (the independent variable); it should be noticed that whereas Y is a random variable once x has been fixed, x is not a random variable but is fixed by and known exactly to the experimenter. The characteristics of the distribution of Y for a given value of x, and in particular $E(Y \mid x)$, are functions of x and may be expected to change with x. The graph of $E(Y \mid x)$ as a function of x is called the *regression* of Y on x; the purpose of regression analysis is to make inferences about the form of this graph.

The simplest and most important type of regression is the straight line

$$E(Y \mid x) = \alpha + \beta x$$

where β is the slope of the line and α its intercept at $x = 0$. As we remarked above, the regression of comb-growth on log dose seems to be approximately linear within the range of doses from $\frac{1}{2}$ mg. to 8 mg.; it cannot, however, be linear over the entire range of doses since (1) there must be an upper limit to the comb-growth as the dose is indefinitely increased, and (2) the comb-growth for zero dose, when log dose $= -\infty$, must be zero and not $-\infty$! This example illustrates the danger of extrapolation.

Let us suppose then that we have n pairs of observations, (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), on two variables of which x is the independent and y the dependent variable and that we wish to estimate the regression of y on x which is assumed to be linear,

$$E(Y \mid x) = \alpha + \beta x.$$

The standard procedure is to choose as the estimated regression line

$$y = a + bx$$

that line which minimises the sum of the squared deviations

of observed from estimated values of y, that is to say the line which minimises the quantity

$$S^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

These deviations are shown graphically in Fig. 31. This method is known as the *method of least squares*. It was first considered in connection with errors of astronomical observations by Legendre in 1806 and by Gauss in 1809.

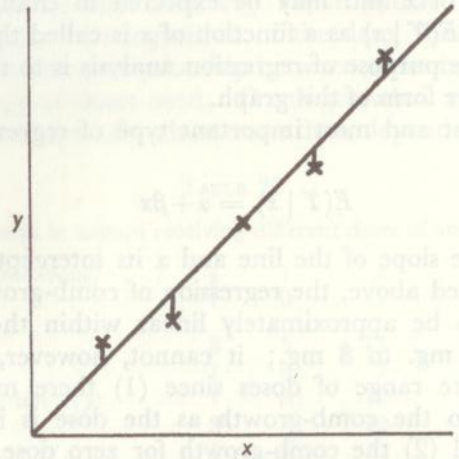

FIG. 31. The 'best' line is the line which minimises the sum of the squares of the deviations in the direction shown

To find the line which minimises S^2 we must solve the pair of simultaneous equations:

$$\frac{\partial S^2}{\partial a} = -2\sum(y_i - a - bx_i) = -2\sum y_i + 2na + 2b\sum x_i = 0$$

$$\frac{\partial S^2}{\partial b} = -2\sum x_i(y_i - a - bx_i) = -2\sum x_i y_i + 2a\sum x_i + 2b\sum x_i^2 = 0.$$

The solution of the first equation is

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

which tells us that the line passes through the point $(\bar{x}, \bar{y})$. Substituting this expression in the second equation we find

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}.$$

It is also convenient to have a formula for S^2, the residual sum of squares which has been minimised. We have

$$\begin{aligned} S^2 &= \sum(y_i - a - bx_i)^2 = \sum[(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \\ &= \sum(y_i - \bar{y})^2 + b^2\sum(x_i - \bar{x})^2 - 2b\sum(y_i - \bar{y})(x_i - \bar{x}) \\ &= \sum(y_i - \bar{y})^2 - b^2\sum(x_i - \bar{x})^2. \end{aligned}$$

The second term in the last expression represents the contribution to the variability of the y's which has been removed by calculating the regression.

We must now consider the justification of this method of estimation. We suppose that the dependent variable is normally distributed with a variance σ^2 which does not depend on x. Our model is then

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where ϵ_i, the random error in the ith observation, is normally distributed with zero mean and variance σ^2. The logarithm of the likelihood of the observations is

$$\log L = -\tfrac{1}{2}n \log 2\pi - \tfrac{1}{2}n \log \sigma^2 - \tfrac{1}{2}\sum(y_i - \alpha - \beta x_i)^2/\sigma^2.$$

Since α and β occur only in the third term, the maximum likelihood estimates of these parameters are found by minimising that term and are thus the same as the least squares estimates. The maximum likelihood estimator of σ^2 is S^2/n. It is also quite easy to show from their sampling distributions (see Appendix) that these three estimators are jointly sufficient for the three parameters of the distribution.

The sampling distributions of a, b and S^2 are investigated in the Appendix to this chapter. It must be remembered that only the dependent variable is considered as a random variable so that these distributions are obtained by imagining that repeated samples of size n are taken with the same, constant values of the x_i's but with different values of the ϵ_i's and hence

of the y_i's. It is shown that a and b are normally distributed, unbiased estimators of α and β respectively with variances given by the formulae:

$$V(a) = \sigma^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i-\bar{x})^2}\right\}$$

$$V(b) = \frac{\sigma^2}{\Sigma(x_i-\bar{x})^2}.$$

These variances are the minimum possible variances of any unbiased estimators of α and β. Furthermore, S^2/σ^2 is distributed as a χ^2 variate with $n-2$ degrees of freedom independently of a and b, so that $s^2 = S^2/(n-2)$ is an unbiased estimator of σ^2. The number of degrees of freedom is 2 less than the number of observations because 2 parameters other than σ^2 have been estimated from the data. The distributions of $\bar{y}$ and b are independent but the distributions of a and b are not independent unless $\bar{x} = 0$, when $a = \bar{y}$. When $\bar{x}$ is positive a and b are negatively correlated, which means that an over-estimate of β is more likely than not to be accompanied by an under-estimate of α, and *vice versa*; when $\bar{x}$ is negative the contrary is true.

These results can be used to perform significance tests or to construct confidence intervals by means of the t distribution. For example,

$$\frac{(b-\beta)}{s}\sqrt{\Sigma(x_i-\bar{x})^2}$$

follows the t distribution with $n-2$ degrees of freedom. This fact can be used either to test a particular value of β, such as $\beta = 0$, which means that there is no relationship between the variables, or to place a confidence interval on β. Inferences about α can be made in a similar way.

It has been assumed that the random errors in the dependent variable are normally distributed with the same variance. This assumption may be wrong in two ways. First, the underlying distribution may not be normal. In this case a and b are no longer normally distributed, but their Expected values and variances are unchanged; S^2/σ^2 no longer follows the χ^2 distribution but its Expected value is still $n-2$. Second, the

variance of Y may not be constant but may depend on x; the regression is then said to be *heteroscedastic* (from the Greek meaning 'different scatter'). In this case a and b are still unbiased estimators and are normally distributed if the underlying distribution is normal, but the formulae for their variances require modification. If the form of the relationship between the variance of Y and x is known, for example if the variance is known to be proportional to x, more efficient estimators can be obtained by weighting the observations with weights inversely proportional to their variances. In general, however, small departures from normality or homoscedasticity will have little effect on inferences about the regression line and may be ignored.

Curvilinear and Multiple Regression

It has been assumed so far that the regression is linear; it is clearly important to be able to test the adequacy of this hypothesis. There are two ways in which this may be done.

Consider the data on the comb-growth of capons in Table 22. We shall change our notation slightly and write y_{ij} for the response of the jth bird at the ith dose level (e.g. $y_{35} = 11$) and $y_{i.}$ for the average response to the ith level (e.g. $y_{2.} = 6{\cdot}2$); note that i and j both run from 1 to 5. If we write the deviation of an observed value, y_{ij}, from its estimated value, $a+bx_i$, in the form

$$y_{ij}-a-bx_i = (y_{i.}-a-bx_i)+(y_{ij}-y_{i.})$$

the residual sum of squares can be split up into two components:

$$S^2 = \sum_i\sum_j (y_{ij}-a-bx_i)^2 = 5\sum_i (y_{i.}-a-bx_i)^2+\sum_i\sum_j (y_{ij}-y_{i.})^2$$

$$= S_1^2 + S_2^2.$$

The factor 5 occurs in S_1^2 because of summation over the index j; the cross-product term vanishes because $\sum_j(y_{ij}-y_{i.})$ is zero for all i.

S_2^2 is the sum of the squared deviations of the observations from their respective means; hence S_2^2/σ^2 is a χ^2 variate