

## 2. Simple Linear Regression

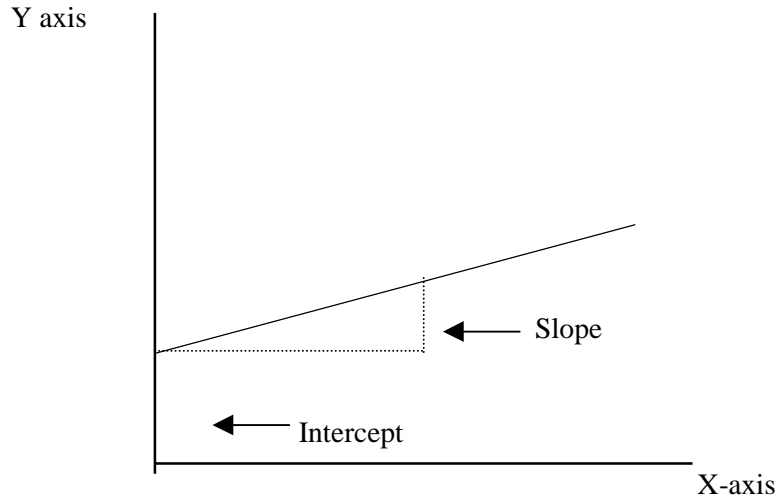
Simple linear regression is a technique in parametric statistics that is commonly used for analyzing mean response of a variable Y which changes according to the magnitude of an intervention variable X. It forms the basis of one of the more important forms of inferential statistical analysis. For example, if we were to perform anthropometric measurements on a group of newborns, we may wish to study how one measurement, say mid-arm circumference, is related to body weight. Such a relationship once precisely established will then enable us to make judgement about body weight of an individual infant if we knew the mid-arm circumference.

### Definition of terms.

In regression analysis there is usually the *independent*, or *explanatory* or *predictor* variable and a *dependent* or *outcome* or *response* variable. (These terms are used interchangeably and are often a source of confusion. Henceforth, we would refer to only explanatory and response variables). The effects of one or more independent variables combine to determine the response variable. In our example of body parameters in the newborn we may decide on mid-arm circumference as the independent and weight as the response variable.

In observational studies the researcher can only observe the independent variable, and use it to predict the outcome variable. For example, in a study of social class and child mortality the observer can only measure different attributes which make up social class (e.g. father's occupation; housing; income; family size; mother's education; and so on) and use the information to predict child mortality. In intervention studies, on the other hand, the researcher is able to manipulate the independent variable (e.g. dose of drug) and thereby predict with greater certainty the outcome. In the first case, it is only possible to identify an association; in the second case a possible causal link.

Relationship between two variables is best observed by means of a *scatter plot*. Then a straight line is drawn which would provide the best estimate of the observed trend. In other words, the line describes the relationship in the best possible manner. Even then for any given value of X there is variability in the values of Y. This is because of the inherent variability between individuals. The line drawn is therefore **the line of means**. In other words, it expresses the mean of all values of Y corresponding to a given value of X.

**Fig. 2.1 The regression Line**

Where the line of means cuts the Y-axis we get *the intercept*. The intercept is the value of Y corresponding to  $X = 0$ . Its units are the units of the Y variable. The line has a *slope*. The slope is a measure of the change in the value of Y corresponding to a unit change in the value of X.

Around the line of means there is variability in the value of Y for any given value of X. This variability is a factor in determining how useful the regression line is for predicting Y for a given value of X.

The above description provides the assumptions on which simple regression analysis is based. This is summarized below:

1. The mean value of the independent variable Y increases or decreases linearly as the value of the independent variable X increases or decreases. To put it simply there is a linear relationship between X and Y.
2. For a given value of the independent variable X the corresponding values of the dependent variable Y are distributed Normally. The mean value of this distribution falls on the regression line.
3. The standard deviation of the values of dependent variable Y at any given value of independent variable X is the same for all values of X. In other words the variability in Y values is the same for all values of X.

4. The deviations of all values of Y from the line of means are independent; i.e. the deviation in any one value of Y has no effect on the other values of Y for any given value of X. The observations are independent, and there is only one pair of observations on each subject.

It is clear that the line of means is an important parameter. Its mathematical representation  $Y = \alpha + \beta X$  is called the **regression equation**, and  $\alpha$  and  $\beta$  are the **regression coefficients**.

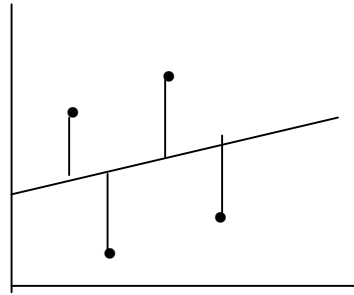
The closer the regression line comes to all the points on the scatter plot the better it is. In other words, one wants a line which would minimize the variation in the observed values of Y for all values of X. Some of the values of Y may well fall on the line. Therefore, minimizing the variation of points around the line is equivalent to "*minimizing the residual variation*". The farther any one point is from the regression line the more the line differs from the data. The values of Y which fall on the regression line (i.e. values which fit the line) are called the **Fits**. The difference between an observed value of Y and a fitted value is the Residual (or **Resid** for short).  $\text{Resid} = \text{Data} - \text{Fit}$ . Each residual may be either negative or positive. Many of the assumptions of linear regression can be stated in terms of the residuals (i.e. observed value of Y – fitted Y). For example:

1. Linear relationship between X and Y can be checked by plotting either Y against X, or residuals against X.
2. The residuals are normally distributed. This can be checked by drawing a histogram or Normality plot.
3. The residuals have the same variability; checked by plotting residuals against fitted values of Y.

The regression line is only part of the description of the relationship between X, Y data points. Being the line of means it tells us only the mean value of Y at any particular value of X. We still need to know how the values of Y are distributed about the regression line. For a given value of X the corresponding values of Y will not only have a mean, but also a variance and a standard deviation. As we saw above, it is one of the basic assumptions of linear regression that this standard deviation is constant, and does not depend upon X.

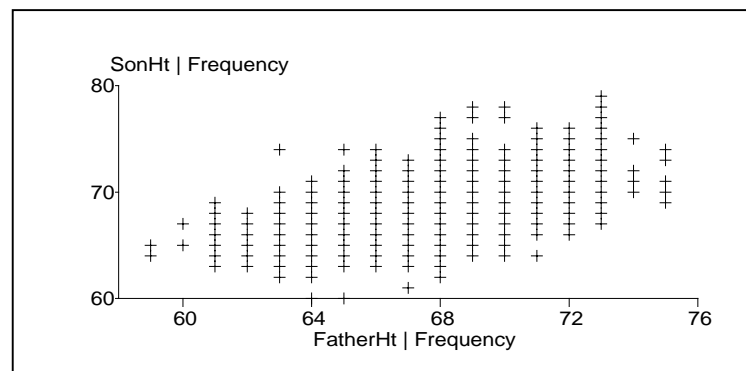
Variation in a population is quantified by the statistical term variance (which is the square of the standard deviation i.e.  $sd^2$ ). This concept is utilized in drawing the line.

The difference between each value of the dependent variable Y not on the line and the value given by the line for a given X (i.e.  $y$ ) is squared and added together. The resulting total is the sum of squared deviations. The smallest value of the sum of squared deviations between the observed values of  $y$  and the regression line is a measure of the "best line". This procedure is called the *method of least squares*, or *least square regression*.

**Fig. 2.2 The Resids**

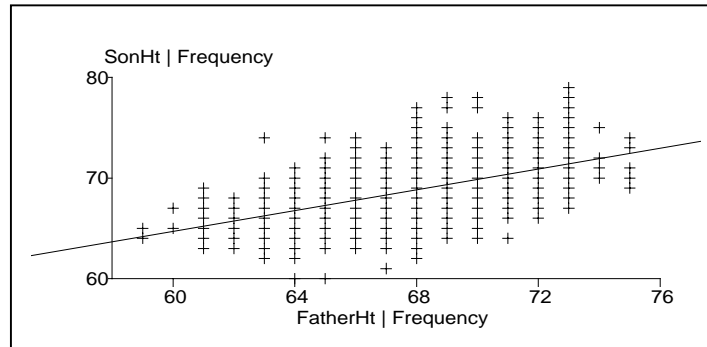
Simple linear regression is the process of fitting a straight line by the method of least squares on a scatter plot to study the relationship between two variables. The intercept and the slope are referred to as the parameters, and the statistical programme finds the values of the parameters that provide the best fit of the regression line.

Let us now see how this works in practice. The heights of 1078 father and son pairs were measured and a scatter plot was drawn as shown in Fig. 2.3.

**Fig. 2.3 Scatterplot of sons' height and fathers' height**

Father-height is plotted on the X-axis, and Son-height on the Y axis. There is considerable scatter for each X value.

We next fit a regression line to the scatter plot as shown below:

**Fig. 2.4 Regression line for the scatterplot in Fig. 2.3**

These two figures illustrate what the regression line of least squares represents. For each individual X value the line represents the mean of the corresponding Y values. Some of the Y values fall on the line. They represent the Fit (or Fitted values). Each Y value away from the regression line would have a residual component, since  $\text{Data} - \text{Fit} = \text{Residual}$ . The concept of the residual is important and we would return to it often in later chapters.

The values of Y that fall on the regression line are the “predicted” or “fitted” values of Y as compared to the observed values. Another way of describing residuals is the difference between the observed and predicted values of Y. The statistical programme minimizes the sum of the squares of the residuals to draw the regression line and find the parameters  $\alpha$  and  $\beta$ . The same applies when one is fitting a curve rather than a straight line.

After the foregoing introduction to the topic of simple linear regression let us now take an example to see how regression works in practice.

#### **A study of breastmilk intake by two methods**

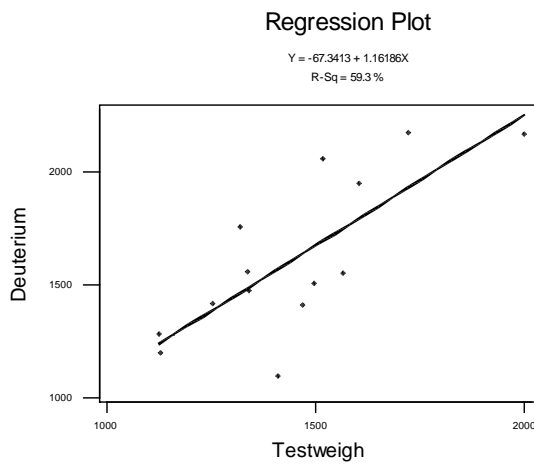
*(Amer. J. Clin. Nutr. 1983; pages 996 - 1003).*

**An experiment was designed to compare two methods of breastmilk intake by infants. The first method was deuterium dilution technique; the second method was test weighing. The researchers selected 14 babies, then used both techniques to measure the amount of milk consumed by each baby. The results are shown below:**

Baby	Deuterium	Test weighing
1	1509	1498
2	1418	1254
3	1561	1336
4	1556	1565
5	2169	2000
6	1760	1318
7	1098	1410
8	1198	1129
9	1479	1342
10	1281	1124
11	1414	1468
12	1954	1604
13	2174	1722
14	2058	1518

We wish to find out as to how well these two methods correlate. We first carry out a scatter plot to see if there is any relationship between the values of breastmilk output obtained by the Deuterium method and by the age old test weighing method.

Fig. 2.5 Scatter plot of Deuterium against Testweighing



A linear relationship is evident, and we can now proceed with performing the regression analysis.

[ In MINITAB → Stat → Regression. In Response box “Deuterium” In Predictors box “Testweigh” ]

The following output is obtained.

(Part I)

The regression equation is  
Deuterium = - 67 + 1.16 Testweigh

(Part II)

Predictor	Coef	StDev	T	P
Constant	-67.3	407.1	-0.17	0.871
Testweigh	1.1619	0.2776	4.19	0.001

s = 234.2      R-Sq = 59.3%      R-Sq(adj) = 56.0%

(Part III)

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	961252	961252	17.52	0.001
Residual Error	12	658387	54866		
Total	13	1619639			

(Part IV)

Unusual Observations

Obs	Testweigh	Deuteriu	Fit	StDev Fit	Residual	St Resid
5	2000	2169.0	2256.4	165.2	-87.4	-0.53 X
7	1410	1098.0	1570.9	63.5	-472.9	-2.10R

R denotes an observation with a large standardized residual  
X denotes an observation whose X value gives it large influence.

### Interpreting the output

(Different portions of the output have been numbered to help with following the explanation provided)

1). The regression equation is

Deut = -67 + 1.16 testweigh which is equivalent to  $Y = -67 + 1.16X$  (Recall that  $Y = \alpha + \beta X$ )

The regression equation represents the least square line drawn through the scatter points. Given the data points it is the best line of least squares that can be drawn to represent them.

2). The next block of output gives the value of  $\alpha = -67.3$  under the heading constant, and  $\beta = 1.1619$  together with additional information.

Next line of output gives ‘s’ the standard deviation which is a measure of how much the observed values of Y differ from the values provided by the regression line. In Fig. 2.5 there is a scatter of data points (n=14) about the regression line. The sum of squares of their vertical distances from the regression line is referred to as **Error Sum of Squares** (SSE for short). The sum amounts to 658387 as shown in the Analysis of Variance table in the next block (Part III) of the output.

$$\begin{aligned} S^2 &= SSE \div n - 2 \\ &= 658387 \div 12 \\ &= 54865.58 \\ s &= \sqrt{54865.58} \\ &= 234.2 \end{aligned}$$

Since s measures the spread of y values about the regression line one would expect 95% of y data points to fall within 2s of that line. The ‘s’ statistic is often called the standard error of the estimate. It plays an important role in answering whether there is evidence to show a linear relationship between X and Y. The programme does this by running a check on  $\beta$ . If there was no linear relationship there will be no regression line, in which case  $\beta = 0$ . To test that  $\beta \neq 0$  the programme divides  $\beta$  by its standard deviation. (Standard deviation of  $\beta$  is derived from s by the formula  $sd \beta = s \div \sqrt{S_{xx}}$ . How  $S_{xx}$  is obtained is described later).

The data in the columns headed stdev and t-ratio are calculated in this manner by the programme, which gives us the standard deviation of  $\alpha$  as 407.1 and of  $\beta$  as 0.2776. The ‘t’ test on  $\beta$  is positive at 0.001, which means that  $\beta \neq 0$ . In other words there is a slope (i.e. there is a linear relationship between x and y data points).

On the same line following ‘s’ is R-sq = 59.3%. R-sq is just the square of the correlation coefficient r. It is the fraction of the variation in Y that is explained by the regression. Next comes R-sq (adj.). It is the value of R-sq adjusted for the sample size (n) and the number of parameters (k) estimated. In this example n = 14 and 2 parameters ( $\alpha$  and  $\beta$ ) have been estimated.

$$\begin{aligned} \text{R-sq(adj.)} &= \text{R-sq} - \{(k - 1) / n - k \times (1 - \text{R-sq})\} \\ &= 0.593 - 1/12 \times (1 - 0.593) \\ &= 0.5590 = 56\%. \end{aligned}$$

3). The next block of information is the Analysis of Variance table. In Analysis of Variance the total variation in Y values is partitioned into its component parts. Because of the linear relationship between Y and X we expect Y to vary as X varies. This variation is called “explained by the regression” or just “regression”. The remaining variation is called “unexplained” or “residual error” or just “error” depending on the programme one uses. Ideally the residual variation should be as small as possible. In that case most of the variation in Y will be explained by the regression, and the points on the scatter plot will lie close to the regression line. The line is then a ‘good fit’.

In the column headed ss the first two entries add up to the last. And so

$ss(\text{Regression}) + ss(\text{Residual}) = ss(\text{Total})$ , which is equivalent to saying  $ss(\text{Regression}) = ss(\text{Total}) - ss(\text{Residual})$ . As explained above, all this means is that  $ss(\text{Regression})$  is the amount of variation in the response variable  $Y$  that is accounted for by the linear relationship between the mean response and the explanatory variable  $X$ .

DF is degree of freedom. SS is sum of squares MS is Mean sum of squares obtained by dividing SS by DF. ( $MS = SS/DF$ ). Total SS is a measure of the variation of  $Y$  about its mean. Here it is 1619639. The Regression SS is the amount of this variation explained by the regression line. (The line of least squares). The output gives it as 961252. Therefore, the fraction of variation explained by the regression line is  $961252 \div 1619639 = 0.593$ . It is more convenient to convert this proportion to percentage which is 59.3% (i.e. the  $R^2$  value), and to say that the regression equation explains 59.3% of the variation in  $Y$ . Thus the Analysis of Variance table helps us to answer a number of questions, like for example,

Is there a straight line relationship between variables  $X$  and  $Y$ , and if so, how strong it is? The  $F$  statistic tests the null hypothesis  $H_0$ : no straight line relationship between  $X$  and  $Y$ . A significant result for the  $F$  statistic means that a relationship exists as described by the straight line model. Prior to that a significant value of the 't' statistic means that the slope  $\beta_1$  is not zero.

4). The last block gives two unusual observations denoted by  $X$  and  $R$ . The explanation is as follows:

Each individual value entered in the regression analysis makes a contribution to the result. But sometimes a particular individual value may stand out from the rest. There are two ways in which the individual value can stand out viz.:

(i). The observed response  $Y$  is very different from the predicted mean response. These are outliers.

(ii). The values of  $X$  are very different from those observed for the other individuals in the sample. These are remote from the rest of the sample since they differ substantially from the other values on the  $X$  axis.

Any value with a standardised residual (obtained by  $\text{Residual} \div \text{stdev Resid}$ ) greater than 2 is labelled as an outlier. Observation 7 thus becomes an outlier by this definition. For example, in row 7 the value of 1098 for "Deuterium" is far lower than what could be expected from "Testweigh" value of 1410. The method of calculating remote from the rest of values is more complex and is not discussed here.

How convincing is the trend observed between the response variable and the explanatory variable?

A regression line is the “best” straight line through a set of data. The intercept and the slope are such that they minimize the sum of squared deviations between the values of the dependent variable  $Y$  at given values of the independent variable  $X$ .

To answer the above question we may ask “Would we be observing a linear trend as strong as that shown if there was no relationship between the dependent and the independent variables?” This question is answered by the  $P$  value of the “ $t$  ratio” of  $\beta$ . If  $\beta$  were to be equal to 0 there would be no slope, which means that the regression line would be horizontal. As we have seen the ‘ $t$  ratio’ is 4.19 and highly significant, which suggests that we have the best available slope for the data.

Secondly, the more tightly the data points are clustered about the regression line the more accurately does the line describe the relationship between the dependent and the independent variables. Close clustering of data points about the regression line means small Residual Sum of Squares ( i.e.RSS; some computer printouts call it Error SS as indeed it is here in the Analysis of Variance table). A small value of RSS means little residual variability about the regression line. But small is a relative term, so how does one judge “small”? It is done by comparing RSS with the sum of squares due to regression (Regression SS in the Analysis of Variance table). The total deviations of all  $Y$  values around the mean of  $Y$  i.e. sum of each individual expression

$(Y_i - \bar{Y})^2$ , in other words  $\sum(Y_i - \bar{Y})^2$  is referred to as Total Sum of Squares. And  $TSS - RSS$  is the sum of squares explained by the regression or ESS. Both RSS (Error SS in the Analysis of Variance table) and Regression SS are first converted into Mean Sums of Squares (**MS**) by dividing with their respective degrees of freedom. In our example, MS for Regression Sum of Squares is 961252, and MS for Error SS is  $658387 \div 12 = 54866$ . Their ratio works out to be  $961252 \div 54866 = 17.52$ . This is called the F ratio. In our example the F ratio is 17.52 and significant. F ratio is an indication of the overall goodness of fit of the regression equation.

### **Further details of simple linear regression**

#### **The statistics of regression and correlation**

As we have seen the regression line gives us the mean value of  $Y$  for any given value of

$X = x$ . The algebraic expression of the regression line is:

$$Y = \alpha + \beta X$$

$\alpha$ , is the intercept. It gives the value of  $Y$  where the regression line meets the  $Y$  axis at  $X = 0$ .

$\beta$ , is the slope, also called the **regression coefficient**. It gives the measure of change in the value of Y for a unit change in the value of X.

There is also a third parameter of interest. For every value of X there is a dispersion of the corresponding values of Y around the regression line. This is the quantity  $Y - (\alpha + \beta x)$ . As we have seen  $\text{Data} - \text{Fit} =$  the residual for each value of Y. It tells us how far the estimated mean Y value is from the actual observed value of Y. This difference between a Y value and the estimated mean Y value is called the **residual**. The main use of residuals is for checking the regression, and we would discuss the procedure a little later. However, in the case of simple linear regression with one explanatory variable the checking has already been performed visually by means of the scatter plot. In situations where there are several explanatory variables the residuals are needed to ascertain whether an additional explanatory variable might be worth considering. To do this the residuals are plotted against the corresponding values of the additional variable. If an association is apparent then the extra variable is a candidate for inclusion. This topic is discussed later in Chapter 5. Since the regression line indicates the mean value of Y there is a standard deviation around this mean. This is called the **residual standard deviation**. It can be calculated from the residual sum of squares as follows:

$$\text{RSS} \div (n - 2) = s^2.$$

which gives  $s = \sqrt{\text{RSS} \div n - 2}$ .

## Correlation

From the given values of X the mean of X ( $\bar{x}$ ) can be calculated, and from this the variance of X which is  $(X - \bar{x})^2$ . The sum of all such calculations for each individual value of X is denoted by the symbol  $S_{xx}$ . Similarly from the given values of Y the mean of Y ( $\bar{y}$ ) can be calculated, and from this the variance of Y, which is denoted by the symbol  $S_{yy}$ . The value of  $\beta$  is obtained by  $\beta = S_{xy} \div S_{xx}$ . One more statistic to consider is  $S_{xy}$ , which is  $\sum(X_i - \bar{x})(Y_i - \bar{y})$ .

The quantity  $S_{xy} \div \sqrt{(S_{xx}S_{yy})}$  is called the correlation coefficient for X and Y, and is denoted by the symbol **r**.

As we have seen in simple linear regression a single outcome variable (assumed to be Normally distributed) is related to a single explanatory variable by a straight line (the regression line) that represents the mean values of the response variable for each value of the explanatory variable. If the line is pointing upwards the relationship is positive, and negative if the direction is downwards. In the former case as the value of X increases so does that of Y. In the latter case as X increases Y decreases. The correlation coefficient is often used as a measure of the strength of association between two variables. A value of **r** close to +1 or -1 indicates a strong linear association. A value close to 0 indicates a weak association. A word of caution. If there is a curved trend the value of **r** could be small

even with a strong relationship. When all the data points fall exactly on the regression line there would be no dispersion around the regression line indicating that the dependent variable can be predicted with certainty from the independent variable. This is the situation where the correlation coefficient is 1. The opposite situation arises when the correlation coefficient is 0. For values in between 0 and 1 the table below is a rough guide:

$r = 0.10$ to $0.29$	Small relationship
or $r = -0.10$ to $-0.29$	
$r = 0.30$ to $0.49$	Medium relationship
or $r = -0.30$ to $-0.49$	
$r = 0.50$ to $1.0$	Large relationship
or $r = -0.50$ to $-1.0$	

The coefficient  $R^2$  (denoted as R-sq in the printout) is a measure of the portion of variability in Y explained by the regression. In other words it is the proportion of variation in Y that can be attributed to X. It is a measure of the strength of the linear relationship between X and Y. Larger the value of  $R^2$  the greater is the linear relationship.  $1-R^2$  is the residual or unexplained variability.

$R^2$  is also referred to as the coefficient of determination. It measures the fit of the regression line to the data. Values of the intercept and regression coefficients are calculated by the programme to minimize residuals in the given data. In all likelihood they will not be identical to the true population values of the intercept and slope. If the fitted regression line is used to predict for the wider situation the prediction will not probably be as accurate as for the sample. A better estimate of the proportion of variability in the population is  $R^2_{Adj}$ . This is discussed in later chapters.

### Analysis of variance

In a regression analysis the variation of Y values around the regression line is a measure of how well the two variables X and Y interact. The method of partitioning the variation is known as the analysis of variance. It is expressed by means of the analysis of variance table.

The total sum of squares  $S_{yy}$  is so called because it represents the total variation of Y values around their mean. It is made up of two parts. The residual sum of squares, and

that explained by the regression line. (Some authors call it reduction due to the regression). In the computer printouts this partitioning is depicted in part (3).

The total variability in the values of Y (the Total Sum of Squares) has been partitioned into that explained by the regression and the part unexplained or Residual variation.

The column headed '**df**' contains the degree of freedom for the different sources of variability. The column headed **MS** contains the mean squares, which are the sum of squares divided by their degree of freedom. The degree of freedom for a variance is the number of observations minus the number of parameters that have been estimated. In a given sample with n observations and an estimate of the mean, the degree of freedom is n-1. In regression analysis the total variability among Y values is the total sum of squares  $S_{yy}$ . The denominator of this variance is n-1 (the degree of freedom). Both of these get partitioned into components associated with regression and residuals so that  $SS_{TOTAL} = SS_{REGRESSION} + SS_{RESIDUAL}$ , as we have seen before, and  $df_{TOTAL} = df_{REGRESSION} + df_{RESIDUAL}$ . Mean squares (MS) are obtained by dividing each sum of squares by the respective degree of freedom.

The F ratio is the ratio  $MS_{REGRESSION} \div MS_{RESIDUAL}$ . A ratio of mean squares follows F distribution and the value of F is reported by the programme. The resulting quantity is then looked up in the table of F distribution (similar to table of *t* and  $\chi^2$ ), and the F - test statistic is calculated to give the value of *P*.