# Programming multiple regression

*By* M. J. R. Healy

**Multiple regression, one form of estimation by least squares, is an important statistical technique. Heavy computing is involved and a great many programs have been written to do the necessary work—so many that it is clear that most of them do not attain an adequate degree of generality. This paper outlines the essential features of regression analysis and attempts to give the essential requirements for a general program.**

## 1. Introduction

*Multiple regression* is the statistician's name for one form of estimation by least squares. It is supposed that the expected value of an observed quantity $y$, called the *dependent variate*, can be expressed as a linear function of a set of quantities $x_i$, $i = 1, \ldots, m$ say, the *independent variates*. If $\hat{y}$ denotes an expected value, we have for a particular observation

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

or, using matrix notation for a whole set of $n$ observations

$$\hat{\boldsymbol{y}} = \boldsymbol{X\beta} \qquad (1)$$

where $X$ has $n$ rows and $m$ columns. (1) is called a *regression equation*, and the $\beta$'s are *regression coefficients*.

The principle of least squares leads us to adopt as estimates of the $\beta$'s the solutions of the *normal equations*

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y} \qquad (2)$$

$(\boldsymbol{X}'\boldsymbol{X})$ is called the *information matrix* of the $x$'s. In many problems, we may assume that the $y$'s are independently distributed about their expected values with a common variance $\sigma^2$. In this case, the variances and covariances of the $b$'s are given by the elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$, and $\sigma^2$ itself can be estimated by

$$s^2 = (\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y})/(n - m). \qquad (3)$$

Multiple regression is an important statistical technique that involves fairly heavy computation, and it is not surprising that most scientific computer installations possess multiple-regression programs—a recent listing of statistical routines (Leone *et al.*, 1961) contains no less than 77 different programs for multiple regression, 32 of of them written for a single type of machine. The inference, borne out by a study of the details of the different programs, is that an adequate degree of generality has not usually been attained, so that new programs have to be written to accomplish tasks for which previous ones made no allowance. The same criticism could be made concerning the programs recently described by Slater (1961). The object of this paper is to outline some of the requirements which may be regarded as essential for a general multiple-regression program.

## 2. Basic computations

The basic computational problems fall into three stages—input and formation of the matrix $X$, formation of the normal equations, and the solution of these equations including the inversion of their coefficient matrix.

### 2.1. *Input*

The design of an input routine for regression calculations has few points that are not shared by other statistical programs—indeed, if a regression program is to be one of a whole set of statistical programs it is desirable that all the input progıams should be as nearly as possible the same, at least so far as the user is concerned.

It is elementary that input should be possible from all the different kinds of peripheral equipment attached to the computer. Assuming this to be so, it should not be necessary for the user of the program to adopt any special format for his data—it should be possible, for example, to accept data assembled either "by rows" (all measurements on one subject together) or "by columns" (all values of one variate together), even though in the latter situation the maximum size of problem that can be tackled will be limited by the necessity of storing essentially all the data.

Facilities should be made available for forming new variates as functions of those read in. The list of built-in functions should include the four arithmetic operations, and a few elementary functions such as the square root, logarithm and exponential. Some more special functions, such as $\log [p/(1 - p)]$ and $\sin^{-1}(p^{\frac{1}{2}})$ with $0 \leqslant p \leqslant 1$, may be included, but it is important to allow the user of the program the possibility of specifying new functions on a temporary basis should the need arise. The notation to be used in specifying these functions calls for some comment. It is often suggested that the rules of some well-known autocode, such as Mercury autocode or FORTRAN, should be adapted to this purpose. It may well be more profitable, in a program that is to be widely used, to adopt a very much more restricted coding of the operations, amounting to no more than a simple 3-address code. The advantage of this is that the simple scheme is quickly learnt and leaves very little room for casual errors in writing the instructions for deriving the new variates. Unless a very complex program is written, a more elaborate scheme is liable to require the observance of so many rules and conventions that infrequent users will make too many mistakes.

When all variate values for a single subject are read in together it is possible to form any required derived values, and to add in the contributions to the information matrix before proceeding to the next subject. Working storage at this stage can thus be limited to that needed for the observations on a single subject, and practically no limitation on the total number of subjects need be imposed. However, if an adequate backing store is available, there are many advantages in storing the data, with or without the derived variate values, for use later in the program (see 2.4 below).

### 2.2. *Formation of the normal equations*

This is the process familiar in statistics as the calculation of sums of squares and products. In equation (2) the elements of $\boldsymbol{X'X}$ and $\boldsymbol{X'y}$ are what the statistician calls *crude* sums of squares and products, and he is accustomed to adjust them to give sums of squares and products of deviations from the means of the variates. This is appropriate whenever the required regression equation contains a constant term. In this case one column of $\boldsymbol{X}$ will consist of 1's, and the corresponding $b$ can be simply obtained from the mean of the $y$'s and the other column means of $\boldsymbol{X}$. Using adjusted sums of squares and products reduces the order of the matrix to be inverted by one, and usually improves its condition in the numerical analyst's sense. The method of making the adjustment requires a little care. The ordinary technique as used on a desk machine is based on the identity

$$\sum_{1}^{n}(x - \bar{x})(y - \bar{y}) = \Sigma xy - \Sigma x \Sigma y/n.$$

Unless $\Sigma x$ and $\Sigma y$ are quite small, this involves the difference of two large numbers, and can lead to a catastrophic loss of significant figures in floating-point arithmetic. As a remedy, the actual deviations from the means may be formed and their products summed, but this involves scanning the matrix of observations twice, which may be undesirable if the matrix is at all large. It is quite sufficient to use deviations from working means fairly close to the true means—the simplest values to use for the purpose are the variate values of the first observation to be read.

A further process may be applied to the matrix before inverting it. It is very desirable from a computational point of view that the elements of the matrix should not differ too much in size. Although a fairly crude scaling is adequate for numerical purposes, the obvious statistical step is to form the matrix of correlations. These quantities will often be of sufficient interest to be worth printing or recording in such a form that they can be used for further analysis if required.

### 2.3. *Matrix inversion and solution of the normal equations*

This topic has been very thoroughly studied in recent years by numerical analysts. The problems arising in multiple regression have certain special features that should be taken into account. The matrix to be inverted is always symmetrical and positive definite, so that the very efficient square-root or Cholesky technique is available; furthermore, elements of the inverse and of the solution vector are seldom needed to more than four significant figures, so that great refinement of numerical technique is not usually necessary.

Loss of numerical accuracy in the course of matrix inversion occurs when the matrix is ill-conditioned. This state of affairs is generally a challenge to the numerical analyst, and he strives to produce accurate inverses of even ill-conditioned matrices. In regression work, ill-conditioning is liable to occur when a pair of independent variates are highly correlated (or become so when the effects of other independent variates are partialled out). The result, which is to produce large elements in the inverse, is to give the regression coefficients large variances and so to render them badly determined, and any effort spent in obtaining numerically accurate values is probably largely wasted. Furthermore, ill-conditioning is very often a sign that the regression problem has been badly posed. A classic example is the regression of the measure of some daily phenomenon on maximum and minimum temperature; this is liable to produce a highly significant regression, yet one in which neither coefficient exceeds its own standard error. If the identical problem is re-posed as a regression on mean temperature and temperature-range, it often turns out that most of the uncertainty is concentrated in the regression on range, the regression on mean temperature being accurately determined. The upshot of this is that the inversion program need not be designed to cope with extreme ill-conditioning, but that it should provide indications of such ill-conditioning as occurs. The simplest of these is the diagonal elements of the triangular square-root matrix. These have in fact a clear statistical interpretation; they are the root-mean-squares of the residuals of the corresponding independent variates after removing the regression on all the previous independent variates. Small values are thus danger signals.

### 2.4. *Output*

The full results of a multiple-regression problem are quite extensive, comprising as they do the vector of regression coefficients, the variance-covariance matrix of these coefficients, and some form of *analysis of variance*—this last might provide the *regression mean square* [$\boldsymbol{b'X'y}/m$ in the notation of equation (2)] and the residual mean square (3), together with their degrees of freedom $m$ and $n - m$ (the contribution from the constant term in the equation is usually excluded from all these quantities). This amount of output is seldom required in full, and parts of it can be provided in a more useful form. The minimum requirement consists of the regression coefficients with their standard errors, and the analysis of variance. Rather than printing the variance-covariance matrix, which is time-consuming and not very illuminating, it should be possible to specify either single linear functions of the regression coefficients, which would then be printed with their standard errors, or sets of such functions, in which case the associated mean square would be printed. If there are $p$ functions with values given by the vector $\boldsymbol{t}$ and variance-covariance matrix $\boldsymbol{C}$, their mean square is $(\boldsymbol{t}'\boldsymbol{C}^{-1}\boldsymbol{t})/p$.

When the regression equation has been estimated, it is possible to calculate the residuals $(y - \hat{y})$ provided the original data are still available. It will seldom be worth printing all of these, but a graphical presentation may be very useful and there are several statistical tests whose results may be presented. Thus a few of the numerically largest residuals may be printed to permit the investigation of aberrant values in the data, and various tests of randomness (see, for example, Durbin and Watson, 1950–51; Stevens, 1939) may be made to guard against failures in the underlying assumptions of the method.

## 3. Exploratory regression

The considerations set out above are those which should govern the simplest type of regression program, where the user is prepared to specify the exact relation that he wishes fitted. In practice, multiple regression is very often used as an exploratory process, several different equations being fitted to the same body of data. A program for this purpose is naturally somewhat more complex.

To begin with, it may well be desirable not to invert the information matrix *ab initio* for each different selection of independent variates—if, for example, it is required to add a set of independent variates to those already in the equation, the existing inverse can be retained and used in a partitioning technique to obtain the new inverse. A good description of the necessary computation is given by Woolf (1951).

The main problem is that of the options to be provided to the user of the program. For a start, he may require identical studies on several different dependent variates; these should normally be run in parallel to avoid repeated inversion of the same matrix. The rest of the specification will consist of a series of steps each ending with the output of a set of regression coefficients with their standard errors and the current analysis of variance.

On the face of it, the user of the program may seem to require the selection of the possible $x$'s which minimizes the residual mean square given by (3), and it is quite possible to write a program which will evaluate (3) for each single $x$, for all pairs of $x$'s, all sets of three $x$'s, and so forth, finally indicating the optimal set. For any appreciable number of possible $x$'s, this is a very lengthy process and one unlikely to be practical except on very fast machines. Moreover, such complete exploration is often neither necessary nor even desirable; when a particular selection of, say, eight $x$'s is needed for a strict minimum of (3), there will usually be a selection of three or four—maybe several such—for which the value of (3) is only slightly increased and which will in fact be of more practical usefulness, while more useful still will be an approach in which the selections of $x$'s investigated are at least to some extent under the user's control.

A possible set of instructions to the program might then be the following.

(1) Fit a specified set of $x$'s.

(2) Fit that one (or pair, or trio) of a specified set of $x$'s that most reduces the residual mean square.

(3) Fit a specified set of $x$'s in the order stated, stopping when one, or two, or three consecutive coefficients all fail to exceed their standard errors by a specified ratio.

(4) From the $x$'s already fitted, remove that one (or pair, or trio) that least reduces the residual mean square.

(5) From the $x$'s already fitted, remove a specified set in the order stated, stopping when the coefficient of the next $x$ to be removed (or the next but one, or the next but two) exceeds its standard error by more than a stated ratio.

Type 1 covers the standard type of regression problem treated earlier; types 3 and 5 are useful when fitting a set of terms that fall into a natural sequence, such as the successive terms in a polynomial. As stated above, each instruction will give rise to a standard output. If further output is required, it will be called for by a special instruction.

## 4. Special problems

There are a number of subsidiary problems in multiple regression, some of which can usefully be catered for by a general program.

### 4.1. *Repeated y-values*

If there are several observations on the dependent variate $y$ for each set of observations on the $x$'s, an independent "within-cell" estimate of $s^2$ can be calculated. This requires the calculation of the mean $y$'s and provides an extra line in the analysis of variance table. It also leads to two further types of instruction for exploratory work.

(6) Fit the following $x$'s in the order stated, stopping when the residual mean square falls below a stated multiple of the within-cell mean square.

(7) From the $x$'s already fitted, remove the following $x$'s in the order stated, stopping when the removal of an $x$ would cause the residual mean square to exceed a stated multiple of the within-cell mean square.

Instructions of types 3 and 5 probably do best as a routine to use the within-cell mean square for estimating standard errors, rather than that derived from deviations from the regression.

### 4.2. *Parallel regressions*

If the observations fall into several groups, it is possible to fit parallel regressions, allowing the coefficient of the constant term to take different values in the different groups but forcing the remaining coefficients to take common values over all the groups [see, for example, Quenouille (1952)]. It is possible to test whether such parallel regressions fit the data as adequately as would regressions fitted to each group independently, and, if so, whether these in their turn could be replaced by merely a single regression.

If there are $p$ groups of observations, $(p + 2)$ regression calculations are required—one for each of the groups, one using within-group sums of squares and products, and one using overall sums of squares and products ignoring the grouping. If the calculation is exploratory, a good deal of storage will be required.

### 4.3. *Weighted regression*

If the original $y$'s differ in accuracy, it may be necessary to take account of this by calculating weighted means and weighted sums of squares and products. Formally, the normal equations (2) become

$$\boldsymbol{X'WXb} = \boldsymbol{X'Wy}$$

where $\boldsymbol{W}$ is a diagonal matrix containing the weights. In practice the weights will be read with the variates, and their application requires only minor changes in that part of the program that forms the normal equations.

### 4.4. *Quantal regression*

In ordinary regression work, $y$ is considered to be a continuous variate. Regression technique is also extremely valuable when applied to so-called quantal data, in which the dependent variate is a proportion. With a dependent variate constrained to lie between 0 and 1 a linear model such as (1) is not very plausible, but this defect can be overcome by transforming the proportions to an unlimited scale. The probit and logit transformations are commonly used for this purpose and the appropriate estimating technique is that of Maximum Likelihood (see Finney, 1952, for details). The necessary calculations involve weighted regression, and because the weights depend upon the expectations the basic procedure has to be used iteratively, but a suitable general program can readily be adapted to this type of calculation.

### 4.5. *Fitting constants to multi-way tables*

As a simple illustration of this procedure, suppose that data are arranged in a two-way table with several items in each cell. If $y_{ijk}$ denotes an observation in the $i$th row and $j$th column, a possible model is

$$\hat{y}_{ijk} = \mu + \alpha_i + \beta_j \qquad (4)$$

where the two factors of the table are assumed not to interact. To estimate $\mu$ and the $\alpha$'s and $\beta$'s by least squares, the normal equations take the form

$$\begin{array}{ll}
n_{..}\mu + n_{1.}\alpha_1 + n_{2.}\alpha_2 + \ldots & \\
 & + n_{.1}\beta_1 + n_{.2}\beta_2 + \ldots = y_{...} \\
n_{1.}\mu + n_{1.}\alpha_1 & + n_{11}\beta_1 + n_{.2}\beta_2 + \ldots = y_{1..} \\
n_{2.}\mu + n_{2.}\alpha_2 & + n_{21}\beta_1 + n_{22}\beta_2 + \ldots = y_{2..} \\
 & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
n_{.1}\mu + n_{11}\alpha_1 + n_{21}\alpha_2 + \ldots + n_{.1}\beta_1 & \qquad\qquad = y_{.1.} \\
 & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot
\end{array}$$

Here $n_{ij}$ denotes the number of observations in the $i$th row and $j$th column, and a dot indicates summation over the corresponding suffix. The model (4) can be extended to more than two factors, and interactions of different orders can be incorporated.

It is a special feature of this type of calculation that the coefficient matrix of the normal equations is singular—in our example, doubly so. To obtain a unique solution, it is necessary to impose two arbitrary linear conditions on the constants. These may be chosen purely for computational convenience; one possibility is to use

$$\begin{aligned} n_{1.}\alpha_1 + n_{2.}\alpha_2 + \ldots &= 0 \\ n_{.1}\beta_1 + n_{.2}\beta_2 + \ldots &= 0 \end{aligned}$$

which gives the solution for $\mu$ immediately, but it is even simpler computationally to take

$$\begin{aligned} \mu &= 0 \\ \beta_1 &= 0. \end{aligned}$$

$\mu$ and $\beta_1$, with their associated equations, can then be merely omitted during the solution process. It will also be seen that the principal submatrices of the coefficient matrix are all diagonal. It is thus easy to eliminate one set of constants from the other equations, reducing (perhaps considerably) the size of the matrix to be inverted.

This technique, though based on the same principles as multiple regression and sharing many of its computational features, is sufficiently different from the user's point of view to make it worth while constructing a special program. This should itself be of considerable generality. Thus, it should allow the user to fit constants by Maximum Likelihood to data in the form of proportions—the technique in this context has proved extremely valuable at Rothamsted but has so far scarcely been exploited elsewhere. It is also useful to be able to impose linear restrictions on the fitted constants. Thus if the categories of one classification can be considered to represent $k$ levels of a quantitative factor $x$, it may be useful to restrain the constants to obey some relation such as

$$\alpha_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2, \qquad i = 1, 2, \ldots, k$$

in which the $\gamma$'s are the quantities of interest.

4.6. *Fitting polynomials*

This is another special case which can quite easily be handled by a general program but which may merit individual treatment. The distinguishing feature is the extreme ill-conditioning that is likely to occur. A satisfactory technique has been described by Forsythe (1957), and its implementation by Clenshaw (1960). A program of this kind can readily be extended to allow the fitting of surfaces in three or more dimensions (Cadwell and Williams, 1961).

## References

Cadwell, J. H., and Williams, D. E. (1961). "Some Orthogonal Methods of Curve and Surface Fitting," *The Computer Journal*, Vol. 4, p. 260.

Clenshaw, C. W. (1960). "Curve Fitting with a Digital Computer," *The Computer Journal*, Vol. 3, p. 170.

Durbin, J., and Watson, G. S. (1950–51). "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, Vol. 37, p. 409, and Vol. 38, p. 159.

Finney, D. J. (1952). *Probit Analysis*, 2nd Edn. Cambridge: University Press.

Forsythe, G. E. (1957). "Generation and Use of Orthogonal Polynomials for Data Fitting with a Digital Computer," *J. Soc Indust. Appl. Math.*, Vol. 5, p. 74.

Leone, F. C., Alanen, J., Andrew, G., and Qureishi, A. S. (1961). *Abstracts of Statistical Computer Routines*, Cleveland: Case Inst. of Technology.

Quenouille, M. H. (1952). *Associated Measurements*. London: Butterworths.

Slater, L. J. (1961). "Regression Analysis," *The Computer Journal*, Vol. 4, p. 287.

Stevens, W. L. (1939). "Distribution of Groups in a Sequence of Alternatives," *Ann. Eugen., Lond.*, Vol. 9, p. 10.

Woolf, B. (1951). "Computation and Interpretation of Multiple Regression," *J. Roy. Statist. Soc. B*, Vol. 13, p. 100.

# Book Review

*Leo and the Managers*, by J. R. M. Simmons, 1962; 174 pages. (London: *Macdonald and Co. (Publishers) Ltd.*, 18s.)

In the first chapter, a series of jerky "flashbacks" take the reader back to 1896 and the Lyons Company's earliest interest in office mechanization. Although the form of presentation is not one which will commend itself to every reader, one is left in no doubt that the company has had a long and continuing interest in office efficiency. The link between LEO (Lyons Electronic Office) and the lessons of office efficiency is made clear. The reader who wishes to learn about the operation of LEO is referred to a long note at the end of the book.

In the succeding three chapters Mr. Simmons develops his "general theory on the organization of business management" and in a final chapter relates his theory to LEO.

The original purpose of his book was to provide "something that could be used by the Central Training Unit of J. Lyons and Company Ltd. to supplement lectures that [he] was then giving to various Company courses on 'The Art and Techniques of Management'." Mr. Simmons has attempted to adapt and expand these lectures, written primarily for Lyons managers, to make them suitable for the general reader interested in the relationship between computers and management.

This creates a difficulty. Mr. Simmons' objective loses its clarity because he is trying to serve two very different audiences at once. The employees of the company should know, for example, when he writes about *actual* company policy and practice and when he is drawing on his imagination to develop his general theory of organization. The general reader cannot know. One cannot help feeling that the general reader would have been better served by an untrammelled description of the Lyons Company organization and the way in which LEO actually assists it to operate effectively. We are, unfortunately, only given tantalizing glimpses of this large organization—for, naturally, all of Mr. Simmons' examples are drawn from it.

A second, and perhaps more fundamental, difficulty arises from the fact that, large as the organization is, it is, as far as one can see, operated as an entity. One thing which we do know about organization theory is that a general theory has not yet been developed from observation of one organization at work. Although such observations may give us some valuable insights into possible relationships, they are unlikely to give us a general theory.

We do, in fact, obtain these insights. Mr. Simmons' approach to the theory of what he describes as "Management Self-Accounting" is refreshing, and in one of the long notes (pages 134–6) the concept is fully described.

In the last chapter he suggests that, given a complex company structure similar to that of Lyons, "It is essential for the best use of a computer for it to be thought of as a means of controlling a decentralized organization and never as an instrument of centralization." To this one may link his final sentence, "If, but only if, the managers are trained to use LEO and they regard it as their own tool, it is capable of being made one of the most powerful management tools that has so far been devised". These thoughts run counter to those who suggest that the future lies in greater centralization. Mr. Simmons suggests that we should continue to push responsibility as far down the "chain of command" as we can. Who knows but that he may be right?

J. H. Leveson