Regression Analysis

By Lucy Joan Slater

This paper sketches the technique of elementary regression analysis, and gives an outline of general *n*-dimensional regression analysis. It then discusses the programming of these calculations for an electronic computer and comments on points that have come out of several years' experience in the practical use of such programs.

Elementary Regressions

Suppose that we are given a set of observations each consisting of two variables, and we wish to know whether there is any underlying connection between one variable and the other. For example, we may record the prices $p_1, p_2, \ldots, p_{12}$ of a commodity at monthly intervals, and the quantities $q_1, q_2, \ldots, q_{12}$ sold during each month, over a period of one year. We can plot the twelve points

$$(p_1, q_1), (p_2, q_2) \ldots (p_{12}, q_{12})$$

on a graph and we find that they lie nearly on a straight line AB (see Fig. 1).

However, we can see that any one of several lines, such as A′B′, might serve equally well, and we find that we have to phrase our question more precisely. In fact, two distinct questions can be asked: how does q vary as p varies? Or, alternatively, how does p vary as q varies?

The "principle of least squares" states that the straight line which answers the first question best is that which gives the least sum of the squares of the distances PC parallel to the q axis, and similarly the straight line which answers best the second question is that which gives the least sum of the squares of the distances PD parallel to the p axis. These two lines do not coincide, unless every observed point falls exactly on the same straight line.

If $q = a + bp$ is the equation of the line, then the sum of the squares of the distances of the twelve points from it is

$$\sum_{m=1}^{12} (q - q_m)^2,$$

where $p = p_m$ and so $q = a + bp_m$, but $q_m \neq a + bp_m$. The condition that this sum should be a minimum in a and b is that the partial derivatives with respect to a and b should be zero; that is

$$\frac{\partial}{\partial a} \Sigma (a + bp_m - q_m)^2 = 0,$$

and

$$\frac{\partial}{\partial b} \Sigma (a + bp_m - q_m)^2 = 0,$$

where the summations are from $m = 1$ to $m = 12 = M$.

But

$$\Sigma (a + bp_m - q_m)^2 = Ma^2 + 2a(b\Sigma p_m - \Sigma q_m) + b^2\Sigma p_m^2 - 2b\Sigma p_m q_m + \Sigma q_m^2$$

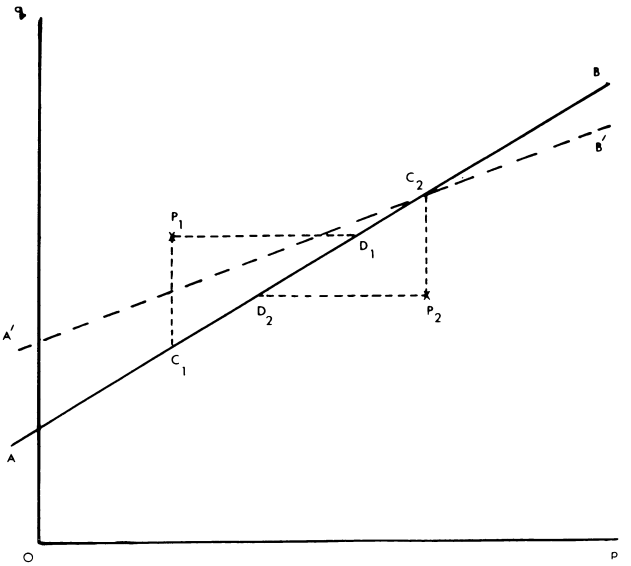

Fig. 1.—Fitting a simple regression line.

and so we arrive at the familiar "normal" equations of elementary statistics,

$$Ma + b\Sigma p_m - \Sigma q_m = 0,$$

$$a\Sigma p_m + b\Sigma p_m^2 - \Sigma p_m q_m = 0,$$

which can be solved to give numerical values for a and for b. Then, when we are given a new price p', we can say that, provided no other influences enter to disturb our market, the quantity we are likely to sell is $q' = a + bp'$.

If we reverse the roles of p and q in the above argument we can produce the other regression line

$$p = a' + b'q$$

to answer the question "Given a new quantity q' what price is it likely to fetch?"

For a full discussion of elementary regression analysis see Brookes and Dick (1955), Croxton and Cowden (1956), Yule and Kendall (1950), or any good statistics book.

Multiple Regression Analysis

In practice we quickly find out that several other factors influence the quantity that we can sell as well as the price which we ask, and so we are forced to seek an

extension of the above simple regression process, to include several variables,

$$x_1, x_2, \ldots, x_n,$$

upon all of which the dependent variable y, the quantity which we are trying to estimate, will depend.

In general the equation of the straight line is replaced by the equation of a hyper-plane

$$y = b_0 + b_1x_1 + \ldots + b_Nx_N,$$

and the sum of the squares of the distances of M observations, each of the form

$$(y_m, x_{m1}, x_{m2}, \ldots, x_{mN}),$$

from this plane is now

$$\sum_{m=1}^{M} (y_m - b_0 - b_1x_{m1} - b_2x_{m2} - \ldots - b_Nx_{mN})^2.$$

The conditions that this should be a minimum are that the partial derivatives with respect to $b_0, b_1, \ldots, b_N$ should all be zero. These conditions lead us to the generalized "normal" equations

$$b_0M + b_1 \Sigma x_{m1} + b_2 \Sigma x_{m2} + \ldots + b_N \Sigma x_{mN} = \Sigma y_m,$$

$$b_0 \Sigma x_{m1} + b_1 \Sigma x_{m1}^2 + b_2 \Sigma x_{m1}x_{m2} + \ldots + b_N \Sigma x_{m1}x_{mN} = \Sigma x_{m1}y_m,$$

$$\ldots\ldots$$

$$b_0 \Sigma x_{mN} + b_1 \Sigma x_{m1}x_{mN} + \ldots + b_N \Sigma x_{mN}^2 = \Sigma x_{mN}y_m,$$

where all the summations are from $m = 1$ to $m = M$.

In the notation of matrix algebra these equations become

$$\begin{bmatrix} N & \Sigma x_1 & \Sigma x_2 & \ldots & \Sigma x_N \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1x_2 & \ldots & \Sigma x_1x_N \\ \Sigma x_2 & \Sigma x_2x_1 & \Sigma x_2^2 & \ldots & \Sigma x_2x_N \\ \ldots & & & & \\ \Sigma x_N & \Sigma x_Nx_1 & \Sigma x_Nx_2 & \ldots & \Sigma x_N^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \ldots \\ b_N \end{bmatrix} = \begin{bmatrix} \Sigma y \\ \Sigma x_1y \\ \Sigma x_2y \\ \ldots \\ \Sigma x_Ny \end{bmatrix}$$

that is $$(X'X)b = X'y,$$

where $X = [x_{ij}]$ is the $M \times N$ matrix of initial observations.

This set of equations has to be solved for the b_i's. The solution is formally

$$b = (X'X)^{-1}X'y,$$

that is

$$\begin{bmatrix} b_0 \\ b_1 \\ \ldots \\ b_N \end{bmatrix} = \begin{bmatrix} c_{00} & c_{01} & \ldots & c_{0N} \\ c_{10} & c_{11} & \ldots & c_{1N} \\ \ldots & & & \\ c_{N0} & c_{N1} & \ldots & c_{NN} \end{bmatrix} \begin{bmatrix} \Sigma y \\ \Sigma x_1y \\ \ldots \\ \Sigma x_Ny \end{bmatrix}$$

where $[c_{ij}] \equiv (X'X)^{-1}$.

For a fuller discussion of the general regression process see Tintner (1952) or Prais and Aitchison (1954).

Associated Statistics

Thus we can see that the central numerical problem in a general regression analysis is the familiar one of inverting a matrix. In this case the matrix is a symmetric one of size $N \times N$. When we have found the values of the b_i's, the statisticians usually call for some statistics to show how good a fit the original observed data is to the calculated regression "line." The most commonly used of these statistics are

$$s^2 = (\Sigma y^2 - b'X'y)/(M - N),$$

$$R_1^2 = b'X'y/\Sigma y^2,$$

and $$R_2^2 = 1 + (b'X'y - \Sigma y^2)/\{\Sigma y^2 - (\Sigma y)^2/M\}.$$

Here s^2 is the measure of the variance, that is an estimate of the dispersion of the errors in the observations, R_2^2 is a measure of the correlation which exists between the observed and the predicted values of y.

In addition we can easily form the complete variance-covariance matrix

$$V(b) = s^2(X'X)^{-1}$$

and we deduce the standard errors $se(b_i)$, of the b_i's. These are the square roots of the diagonal elements of $V(b)$, that is

$$se(b_i) = \sqrt{(s^2c_{ii})}.$$

Then from the vectors $\{b_i\}$ and $\{se(b_i)\}$ can be deduced the vector $\{t_i\}$, where $t_i = b_i/se(b_i)$, and we can apply some simplified tests of significance to the t_i's to determine how "normal" or otherwise the results are. For about 5% of the results to be significant we test $|t_i| \geqslant 2$, and regard the 95% of the results for which $|t_i| < 2$ as not significant.

Finally we may be asked to calculate the residuals, that is the M differences e_m between the observed values y_m and the calculated values $\sum_{n=0}^{N} b_nx_{mn}$, and as a measure of the dispersion of these residuals, the Durban–Watson d-statistic defined as

$$d = \sum_{m=2}^{M} (e_m - e_{m-1})^2 / \sum_{m=1}^{M} e_m^2.$$

If the numerical value of d is near 2, it indicates that the distribution of the observations is "normal," with the observed points scattered more or less at random equally on both sides of the regression line. But if the numerical value of d is not near 2, it indicates that the distribution of errors in the observations is not a normal one and that the regression "line" is likely to give a poor fit to the observed data (Watson, 1951).

Two further points arise. Firstly, not all regressions follow the linear law. It is quite usual for some simple functions to be introduced so that the data are in fact fitted to a curve rather than to a straight line. For example, the usual law for demand analysis is $y = a + b \log_e x$, so that we might transform the observed x's into $x' = \log_e x$ before fitting a straight line $y = a + bx'$ to our data.

Secondly, many regressions are of the type known as weighted. We may feel that some of the observations are of greater importance than others, and so we attach a weight w_m to each observation. The regression matrix is then built up as

$$X'X = [\Sigma w x_i x_j],$$

the vector becomes

$$X'y = \{\Sigma w x_i y\},$$

and the three main statistics become

$$s^2 = (\Sigma w y^2 - b'X'y)/(M - N),$$

$$R_1^2 = b'X'y/\Sigma w y^2,$$

and $$R_2^2 = 1 + (b'X'y - \Sigma w y^2)/\{\Sigma w y^2 - (\Sigma w y)^2/\Sigma w\}.$$

It can now be seen that the unweighted calculation is exactly the same as the weighted calculation with every weight treated as unity, so that one computer program can do both types of calculation.

Early Programs

Now how do we organize all these computations for an electronic computer? First we require an accurate, fast, and efficient matrix-inversion routine, which takes up as little space as possible in the store. The Jourdan process described by Barron and Swinnerton-Dyer (1960) is the standard routine used in the Cambridge University Mathematical Laboratory. This builds up the inverse, by the usual process of triangulation and back substitution, on top of the original matrix in the store, and thus it requires only one extra row of storage space, so that an $n \times n$ matrix can be replaced by its inverse, using only $n^2 + n$ storage locations.

Secondly, we need to organize the whole calculation to use as little storage space as possible, storing our set of data, if it is large, on auxiliary storage such as magnetic tape. Since the operation of the program does not take much machine time, and the time used is occupied mainly in reading in the data and putting out the results, it is well worth while to spend some time and trouble in finding out exactly which of the many possible statistics calculated will really be read by the statistician who will receive the results! Usually he is only interested in $\{b_i\}$, $\{se(b_i)\}$, s^2, R_1^2, and R_2^2 unless these results show a high degree of significance, when the other items should be printed.

A series of programs has been developed for the Department of Applied Economics, Cambridge University, over the past five years, and these have been used successfully on a wide variety of problems, not all economic ones. The first programs were of a very simple type. The data were punched in rows, in the order

$$y_1, x_{11}, x_{12}, \ldots, x_{1N},$$
$$y_2, x_{21}, x_{22}, \ldots, x_{2N},$$
$$\ldots\ldots$$

and read in one row at a time. The matrix

$$X'X = [\Sigma x_i x_j]$$

and the vector

$$X'y = \{\Sigma x_i y\}$$

were built up in the machine, together with Σy and Σy^2. Then the matrix $(X'X)^{-1}$ was calculated and the vector

$$b = (X'X)^{-1}X'y$$

was printed together with the vector $se(b)$, and the three statistics

$$s^2,\ R_1^2, \text{ and } R_2^2.$$

An improved version of this program made some elementary transformations of the data during input. These transformations were controlled by a code number at the head of each data tape, so that, for example, 13 25 would mean that the first three variables were to be treated plain, and the next five variables were to have logarithms formed. Provision was made for weighted data, by punching the data in rows, as

$$w_m, y_m, x_{1m}, x_{2m}, \ldots, x_{Nm}$$

and forming the matrix

$$[\Sigma w_m x_{im} x_{jm}]$$

and the vector $\{\Sigma w_m x_{im} y_m\}$ as outlined above. Here w_m was punched as unity if no weighting was required.

Residues were then formed by re-reading the data tape and reforming the data in the machine.

Some Points of Difficulty

This program was used for several years and the following points were found to be of importance. First, the complaint was made that s^2 was not very accurate, and that double-length working should be used to improve its accuracy. This was tried but with no improvement, and if we pause to think we can easily see why. If we put three significant figures into our initial data, and form the sums $\Sigma x_{im} x_{jm}$ accurately, we may gain one significant figure with luck, giving four in all. When we form Σy^2 and $b'X'y$, they are by definition likely to be nearly equal. If they are equal, to three significant figures, as they frequently are in practice, no double-length working will enable us to find any more accuracy than the one significant figure that we have inherited from the first part of the calculation. This is a point which is obvious to any experienced computer on desk machines, but one which theoretical workers find very hard to grasp.

Second, some trouble was caused by the occasional appearance of a negative s^2, since it is necessary to form $\sqrt{(s^2 c_{ii})}$ in order to calculate $se(b)$. This negative sign arose either when s^2 was very small, that is when $b'X'y$ was equal to Σy^2, to the number of figures that were significant, and the "mess" on the end of the accumulator then appeared as a very small negative number,

or if c_{ii} became negative during the inversion process, for a similar reason. Since s^2, by definition, must always be positive, to have any meaning at all, as a first palliative measure, to prevent the calculation from coming to an abrupt halt, a modulus sign was added to the standard error formula, so that $se(b_i) = \sqrt{|s^2 c_{ii}|}$.

Also, several tests were inserted at various points in the programs to guard against some of the possible errors in the data tapes. For example, check that $M - N > 0$ since s^2 has no meaning if $M = N$, or, check that the number of items read was actually $M(N + 1)$, the size of the initial data.

The first trouble was completely overcome when provision was made to remove the means from the actual data, *before* building it into the $X'X$ matrix, instead of removing the means from the $X'X$ matrix itself, as the first program had done. The second trouble was also alleviated when the means were removed from the data, as above, but its occasional reappearance was not completely cured until the much improved matrix inversion routine (Barron and Swinnerton-Dyer, 1960) came into use.

The third point (which only came out after some practical use of the program) was that, after the original data had been checked in every possible way, much of the clerical effort required to use the program went into the preparation and re-ordering of data into the correct form, and the making up and correction of data tapes. The data for regression analysis usually comes as a sheet of numbers, which might be treated as in rows or in columns, accompanied by the equation required in the regression. This may seem at first sight to be anything but linear in form, for example

$$\log_e y = a + b \log_e{}^2 x_1/x_2 + c \exp(x_3 - x_4).$$

There would also be a laconic instruction "col. 1 is y, take cols. 2–9 as x_1, x_2 in turn, and cols. 10–17 as x_3, x_4." The equation is reduced to

$$y' = a + bx_1' + cx_2',$$

where $y' = \log_e y$,
$x_1' = \log_e{}^2 x_1/x_2$,
and $x_2' = \exp(x_3 - x_4)$.

It was then necessary to make these transformations of the data outside the machine, on desk calculators, and to copy out, prepare, punch, and check eight different long data tapes. When we are satisfied with the initial data, it cannot be said too strongly that, if at all possible, no operations on this initial data should be performed outside the machine, except the actual punching of the data tapes.

A General Program

The latest version of the program devotes some space to reading in control tapes which can transform the same set of data in many different ways. The main data is read in as a whole, preceded by a code number to instruct the machine to treat it as punched in rows or columns. Then the control tape is read; this consists of groups of code numbers, to select and transform the different variables, and then to build up the matrix $X'X$, as usual. The main data is held in the store, ready to be re-interpreted for the calculation of the residuals if required, or retransformed for the trial of another regression equation. By this means the initial data tapes are both fewer in number and very much shorter in length.

In practice it has been found that a set of data rarely exceeds 15×30 and that the matrix required is seldom greater than 10×10.

Thus EDSAC 2, with 1,024 storage registers, has plenty of room for the program, the data, and the matrix.

The program first reads in and stores the main data preceded by three code numbers, 0 or 1, h and M. If the first number is 0 the data is treated as having been punched in rows, if it is 1, as having been punched in columns; if it is neither 0 nor 1 the program prints a special symbol to indicate a faulty data tape.

The program next reads one control tape, and checks that its layout is correct. The layout of one of these control tapes is

0 or 1	for do not print or print residues,
p	where p% is the level for significant t_i's,
N	where N is the number of x's in this regression, including a constant as x_0 if required,
0 or 1	for do not remove means or remove means,
$N + 2$	groups of four digits each, for w_m, y_m, and the N x_{nm}'s.

At the end of one such control tape, the third part of the program builds up the selected data from the main data, and transforms it, ready to be used in the particular regression called for, according to each of the $N + 2$ groups of digits in turn.

The first digit d_1 of each group selects an element z_1 from the row of main data being treated, and the second digit d_2 selects another element z_2 from the same row. The third digit d_3 can be 0, 1, . . ., 11, 12, and it causes the formation of $f(z_1, z_2) = z_1'$ where

0 is the null transformation $z_1' = z_1$,
1 is $z_1' = z_1 + z_2$,
2 is $z_1' = z_1 - z_2$,
3 is $z_1' = z_1 \times z_2$,
4 is $z_1' = z_1/z_2$,
5 is $z_1' = \log_e z_1$,
6 is $z_1' = \exp z_1$,
7 is $z_1' = z_1^2$,
8 is $z_1' = 1/z_1$,
9 is $z_1' = \log_e^2 z_1$,
10 is the difference $z_1' = \Delta z_1 = z_{1m} - z_{1m-1}$,
11 is $z_1' \equiv 1$,
12 is $z_1' = 1/z_1^2$.

The fourth digit d_4 has a similar significance, except that now the transformation is performed on z_1' and z_2'

instead of on z_1 and z_2. Other transformations can easily be inserted in the program at this point, using higher digits, 13, 14, . . . to indicate them.

When one row of data has been built up as

$$w_m, y_m, x_{m1}, x_{m2}, \ldots, x_{mN},$$

where $w_m = 1$ for plain regression (and the means

$$\frac{\Sigma wy}{\Sigma w}, \quad \frac{\Sigma wx}{\Sigma w}$$

have been removed if required), this data is accumulated into the symmetric matrix $X'X$, the vector $X'y$, and the quantities Σw_m, $\Sigma w_m y_m$, $\Sigma w_m y_m^2$. The program cycles over M rows of data, and then having completed the formation of $X'X$, it inverts $X'X$. Then the vector $\{b_i\}$ is calculated as

$$(X'X)^{-1}X'y = \left\{ \sum_{j=1}^{N} c_{ij} \sum_{m=1}^{M} w_m x_{mi} y_m \right\}$$

where $[c_{ij}] = (X'X)^{-1}$, and printed, together with the three statistics s^2, R_1^2, and R_2^2. Every element of the inverse matrix $[c_{ij}]$ is multiplied by s^2, to form the variance-covariance matrix $V(b)$ in the store. This is checked for symmetry, and the standard errors are formed as the square roots of the diagonal elements of $V(b)$, that is $se(b_i) = \sqrt{|(s^2 c_{ii})|}$, and printed. The vector $\{t_i\} = \{b_i/se(b_i)\}$ is calculated and each element is tested in turn for significance. If any $|t_i| \geqslant p$, then the whole vector $\{t_i\}$, is printed together with M, N, Σw_m, $\Sigma w_m y_m$, $\Sigma w_m y_m^2$, $X'y$, and the matrix in triangular form $V(b)$.

If no $|t_i| \geqslant p$, then all this printing is omitted. If the residues are called for, the data is then rebuilt according to the control numbers, from the original data, which is still held in the store, and the residues are formed from

$$e_m = y_m - \sum_{i=1}^{N} b_i x_{im}.$$

The program again cycles over the M rows of data.

Finally

$$d = \sum_{m=2}^{M} (\Delta e_m)^2 / \sum_{m=1}^{M} e_m^2,$$

is printed, and the program resets itself ready to read another control tape, or a special symbol at the end of the sequence of control tapes. This symbol resets the program in another way, so that it is then ready to read in a new set of main data.

Acknowledgements

Thanks are due to Dr. M. V. Wilkes, the Director of the Cambridge University Mathematical Laboratory, for permitting the use of EDSAC 2 for the experiments reported in this paper, and also to J. A. C. Brown of the Department of Applied Economics, Cambridge University, for many useful discussions.

References

BARRON, D., and SWINNERTON-DYER, H. P. F. (1960). "Solution of Simultaneous Linear Equations using a Magnetic-Tape Store," *The Computer Journal*, Vol. 3, p. 28.

BROOKES, B. C., and DICK, W. F. L. (1955). *Introduction to Statistical Methods*, London: W. H. Heinemann Ltd.

CROXTON, F. E., and COWDEN, D. J. (1956). *Applied General Statistics*, London: I. Pitman & Sons Ltd.

PRAIS, S. J., and AITCHISON, J. (1954). "The Grouping of Observations in Regression Analysis," *Review of the International Statistical Institute*, Vol. XXII, p. 1.

TINTNER, G. (1952). *Econometrics*, New York: J. Wiley & Sons.

WATSON, G. S. (1951). *Serial Correlation in Regression Analysis*, Institute of Statistics, Mimeograph No. 49, North Carolina University.

YULE, G. U., and KENDALL, M. G. (1950). *An Introduction to the Theory of Statistics*, London: Griffin.