Simultaneous Regression Equations in Experimentation

Prepared Under Contract No. DA-36-034-ORD-1517 (RD)

(Experimental Designs for Industrial Research)

by

E. J. Williams

Institute of Statistics

Mimeo Series No. 173

April, 1957

Simultaneous Regression Equations in Experimentation

by

E. J. Williams

1. Introduction

The purpose of this paper is to discuss the determination and interpretation of simultaneous equations fitted to experimental data. Although little has been written on simultaneous equations in experimentation, their uses in economics have frequently been discussed. In that field, however, there is often no distinction between dependent and independent variables. In what is known in econometrics as a complete system of simultaneous equations, there are as many equations as endogenous variables so that the equations consist of a linear transformation from the unknown disturbances and known exogenous variables to the observed variables. The treatment of simultaneous equations in econometrics is generally troublesome and depends on the completeness of the system of equations, and the identifiability of the parameters.

In experimental work, on the other hand, there is in many situations a clear distinction between the dependent and independent variables. Thus the number of equations will be at most equal to the number of dependent variables. In this field, too, there is a case of particular interest, as will be shown below, which occurs when the numbers of dependent and independent variables are the same. The applications of simultaneous equations to experimental work seem to be quite important and are much more straightforward than those in econometrics, yet, strangely enough, they seem to have been little discussed. The only published work in this field with which we are familiar is that of Box and Hunter (1954), but even this relates to a different situation from that considered here, and to particular applications in experimental design.

We begin by discussing a simple application of simultaneous equations to experimental work. Then will follow the mathematical theory, after which special cases will be discussed. It seems that the relatively early introduction of the linear discriminant function has diverted the attention of statisticians from simultaneous equations; it appears that in many cases which are dealt with by discriminant functions, the set of simultaneous equations (from which the discriminant function can be derived) is more informative.

2. A chemical example

Fisher, Hansen and Norton (1955) discuss the quantitative determination of glucose and galactose simultaneously in solutions of unknown chemical composition, by means of optical density measurements. Without going into the technical details, which are given in the paper referred to, we can simply state that solutions of glucose and galactose are treated to develop a color; the optical density of the solution to light of two different wavelengths is then determined, and the two data thus obtained are used to estimate the amount of each sugar in solution. It is assumed that, within the range of concentrations studied, optical density for each sugar is proportional to amount of sugar; then use is made of the fact that each sugar differs in its density to light of different wavelengths.

Solutions containing known amounts of glucose and galactose were prepared, and the density at two different wavelengths (470 and 560 mμ) determined. The data enable a regression of density on amount of each sugar to be determined for each wavelength. These regressions then constitute a calibration of the apparatus, such that if optical densities for some unknown solution are substituted in the equations, the amount of each sugar can be estimated.

Thus, if y_1 and y_2 are the optical densities at 470 and 560 mμ respectively,

and x_1 and x_2 the amounts of each sugar (in milligrams), the regression equations may be written

$$\begin{aligned} Y_1 &= b_{11}x_1 + b_{21}x_2 \\ Y_2 &= b_{12}x_1 + b_{22}x_2 \end{aligned} \tag{1}$$

These equations have no constant term, since the optical densities are zero at zero concentration of the sugars. In the practical use of these equations, the y's will be observed values and the x's predicted. If the equations are solved for this purpose, we get

$$\begin{aligned} X_1 &= b^{11}y_1 + b^{21}y_2 \\ X_2 &= b^{12}y_1 + b^{22}y_2 \end{aligned} \tag{2}$$

where the matrix

$$\begin{bmatrix} b^{11} & b^{21} \\ b^{12} & b^{22} \end{bmatrix}$$

is the inverse of the original matrix of regression coefficients,

$$\begin{bmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \end{bmatrix}$$

The equations (2) will be called inverse regression equations, and the X-value inverse estimates. It will be seen that in practically every calibration problem, inverse estimates are required, since the quantities arbitrarily assigned in the calibration are unknown in the application to estimation.

Problems of this kind must be of frequent occurrence in quantitative chemical analysis and in other fields. The determination of the accuracy with which estimates can be made from such equations is an important practical problem. We now give the mathematical derivation of sampling errors and fiducial intervals, before

returning to the arithmetical analysis of the example just discussed.

3. Simultaneous equations in general

In general, we may consider that we have n observations on each of p independent variables x_i ($i = 1, 2, \ldots\ldots, p$) and q dependent variables y_j ($j = 1, 2, \ldots\ldots, q$), and that we require to estimate the y_j in terms of the x_i or vice versa. Then we may determine q regression equations

$$Y_j = \sum_i b_{ij} x_i \qquad (j = 1, 2, \ldots\ldots, q) \qquad (3)$$

in which for simplicity the variables are measured from their means so that the constant terms vanish.

We adopt the following notation:

t_{hi} sum of products of x_h and x_i (n-1 degrees of freedom)

u_{jk} total sum of products of y_j and y_k (n-1 degrees of freedom)

v_{jk} residual sum of products of y_j and y_k (n-p-1 degrees of freedom)

$T = (t_{hi})$ $\qquad T^{-1} = (t^{hi})$

$U = (u_{jk})$

$V = (v_{jk})$ $\qquad V^{-1} = (v^{jk})$

Lower case x_i or y_j will denote either observed or potentially observed (though sometimes actually unknown) quantities, while capital X_i or Y_j will denote estimates based on the observed quantities.

4. Direct estimation

If one of the y_j or a linear combination of them is to be estimated from the equations (3), the procedure is straightforward. For the variances of the regression coefficients we have the familiar results

$$(n-p-1)V(b_{ij}) = v_{jj}t^{ii}$$

and generally

$$(n-p-1)\mathrm{Cov}(b_{hj},b_{ik}) = v_{jk}t^{hi}$$

$$= (n-p-1)\mathrm{Cov}(b_{ij},b_{hk}) \qquad (4)$$

Hence, for the variance of an estimate, we have

$$(n-p-1)V(Y_j) = v_{jj}\left(\frac{1}{n} + \sum_h\sum_i t^{hi}x_h x_i\right) \qquad (5)$$

and in general for the covariance of any two estimates,

$$(n-p-1)\mathrm{Cov}(Y_j,Y_k) = v_{jk}\left(\frac{1}{n} + \sum_h\sum_i t^{hi}x_h x_i\right), \qquad (6)$$

the term $\frac{1}{n}$ being included to allow for the fact that the variables are measured from their means.

In order to know how much a new observation y_j will vary about the predicted value, we need the variance about an estimate, as well as the variance of the estimate. We have

$$(n-p-1)V(y_j-Y_j) = Hv_{jj} \qquad (7)$$

and generally,

$$(n-p-1)\mathrm{Cov}(y_j-Y_j,y_k-Y_k) = Hv_{jk} \qquad (8)$$

where

$$H = 1 + \frac{1}{n} + \sum_h\sum_i t^{hi}x_h x_i \qquad (9)$$

If it is required to estimate a linear combination of the y_j, for example

$$y_a = \sum_j a_j y_j$$

the regression coefficients are linear combinations of the original coefficients, viz.,

$$b_{ia} = \sum_j a_j b_{ij} \quad ,$$

and the regression equation for estimating y_a may be written

$$Y_a = \sum_i b_{ia} x_i \quad .$$

The variance of an estimate is given by

$$(n-p-1)V(Y_a) = \sum_j \sum_k a_j a_k v_{jk} \left(\frac{1}{n} + \sum_h \sum_i t^{hi} x_h x_i\right) \qquad (10)$$

A special case of a linear combination of the dependent variables is Hotelling's "most predictable criterion". For a linear combination with coefficients a_j, the residual sum of squares after fitting the regression on the x_i is

$$\sum_j \sum_k a_j a_k v_{jk} \qquad (11)$$

and the total sum of squares is

$$\sum_j \sum_k a_j a_k u_{jk} \qquad (12)$$

The linear combination which minimizes the ratio of (11) to (12) will clearly be an estimate of that linear combination which is least affected by departure from regression, and has been designated by Hotelling the most predictable criterion. The coefficients a_j will be found as one of the latent vectors of the matrix $V^{-1}U$. Whether this linear combination has any relevance to the interpretation of the data will depend on the nature of the problem.

5. Inverse estimation

As mentioned earlier, we are most often interested in using a set of simultaneous regression equations inversely for estimating values of the independent variables from observed values of the dependent variables. This situation arises frequently, for example, in calibration experiments, as the above discussed example shows. Now in order that the regression equations may be solved for the independent variables, it is necessary, generally speaking, that the number of equations equal the number of independent variables. If there are fewer equations than independent

variables, they cannot be solved, and all that can be determined are certain relationships among the estimated values of the independent variables. On the other hand, if there are more equations than unknown independent variables, we have redundant information; however, by an adaptation of the method of least squares, valid estimates of the unknowns may be determined. In this case the discrepancies of the individual equations from these estimates provide a measure of the consistency of the different equations and hence of the different dependent variables. We shall consider each of these cases in turn.

(a) $p = q$

In this case the regression equations (3) may be solved directly to give the estimates of the x_i, which we denote, without risk of confusion with direct estimates, by X_i. The solutions are

$$X_i = \sum_j b^{ji} y_j, \tag{13}$$

where the b^{ji} are the elements of the matrix inverse to the square matrix

$$B = (b_{ij})$$

We note that in the matrix B, rows correspond to x-variables and columns to y-variables, while in B^{-1}, rows correspond to y-variables and columns to x-variables. Thus, for either direct or inverse regression equations, the regression coefficients corresponding to any predictand are read down the columns.

We shall show below how tolerance limits for values, corresponding to the estimates X_i, may be determined by means of the F-test. First of all, however, it is of interest to determine approximate standard errors for these estimates. These standard errors will be applicable when the estimated regression coefficients are large compared with their standard errors, and the inverse regression coefficients are likewise large compared with their standard errors. This second condition

requiree in particular that the matrix B be not almost singular.

Now we have

$$BB^{-1} = I \quad ;$$

hence, on taking differentials, and multiplying the results by B^{-1}, we find

$$dB^{-1} = -B^{-1}(dB)B^{-1} \tag{14}$$

whence

$$db^{ji} = -\sum_{h}\sum_{k} b^{jh}b^{ki}db_{hk} \tag{15}$$

The equations (14) and (15) represent a linear transformation of the differentials db_{hk}. Taking the direct product (van der Waerden, 1931) of such a transformation and its transposed, we have

$$\begin{aligned} dB^{-1} \underline{x}\, dB'^{-1} &= (B^{-1}(dB)B^{-1}) \underline{x} (B'^{-1}(dB')B'^{-1}) \\ &= (B^{-1} \underline{x} B'^{-1})\ (dB \underline{x}\, dB')\ (B^{-1} \underline{x} B'^{-1}) \end{aligned} \tag{16}$$

Now each of the direct products in equation (16) is a $p^2 \times p^2$ matrix whose typical elements are products of two regression coefficients or differentials. For instance the typical element of $dB \underline{x}\, dB'$ is

$$db_{ij}db_{i'j'} \quad .$$

If we take expectations of each side of equation (16), we get on the left hand side the matrix of variances and covariances of the b^{ji}, while on the right hand side the middle factor gives the matrix of variances and covariances of the b_{ij}. Now, as we have seen, the appropriate estimate of the covariance of b_{ij} and $b_{i'j'}$ is

$$t^{ii'}v_{jj'}/(n-p-1) \quad .$$

If we make a suitable permutation of rows and columns, the expected value of the middle factor therefore becomes the direct product

$$T^{-1} \underline{x}\, V/(n-p-1) \quad ,$$

of two p x p matrices.

Denoting the estimated expected value of the left hand side by W, we find, again after suitable permutations of rows and columns, that

$$(n-p-1)\,W = (B^{-1} \underline{x} B'^{-1})\,(T^{-1} \underline{x} V)\,(B'^{-1} \underline{x} B^{-1})$$

$$= (B^{-1}T^{-1}B'^{-1}) \underline{x} (B'^{-1}VB^{-1}) \qquad (17)$$

$$= M^{-1} \underline{x} Q^{-1}$$

where $M = B'TB$

and $Q = BV^{-1}B'$

so that $M^{-1} = B^{-1}T^{-1}B'^{-1}$

and $Q^{-1} = B'^{-1}VB^{-1}$

This result gives in particular

$$(n-p-1)\,\mathrm{Cov}\,(b^{ji},b^{j'i'}) = m^{jj'}q^{ii'}$$

$$= \sum_{h}\sum_{h'} t^{hh'}b^{jh}b^{j'h'} \sum_{k}\sum_{k'} v_{kk'}b^{ki}b^{k'i'}$$

$$= (n-p-1)\,\mathrm{Cov}\,(b^{ji'},b^{j'i}) \quad . \qquad (18)$$

These results are, of course, approximate and will often be inaccurate; their interest lies in the fact that the expressions found are similar to those occurring in the exact analysis.

We may now determine approximate variances and covariances of estimates X_i based on observations y_j.

$$(n-p-1)V(X_i) = (n-p-1)V\Big(\sum_j b^{ji}y_j\Big)$$

$$= (1+\tfrac{1}{n})\sum_{j}\sum_{j'} v_{jj'}b^{ji}b^{j'i} + \sum_{j}\sum_{j'}\sum_{h}\sum_{h'} y_j y_{j'} t^{hh'}b^{jh}b^{j'h'} \sum_{k}\sum_{k'} v_{kk'}b^{ki}b^{k'i}$$

$$= \sum_{j}\sum_{j'} v_{jj'}b^{ji}b^{j'i}\,(1+\tfrac{1}{n}+\sum_{h}\sum_{h'} t^{hh'}X_h X_{h'}) \qquad (19)$$

This result follows from the formula for the approximate variance of a product, and from the fact that the b^{ji} and the y_j are independent. Similarly, to the same degree

of approximation,

$$(n-p-1)\ \mathrm{Cov}\ (X_i,X_{i'}) = \sum_{j}\sum_{j'} v_{jj'}b^{ji}b^{j'i'}\ (1+\frac{1}{n}+\sum_{h}\sum_{h'} t^{hh'}X_hX_{h'}) \qquad (20)$$

The covariance matrix of the X_i may be written $\frac{HQ^{-1}}{n-p-1}$ where $Q^{-1} = B'^{-1}VB^{-1}$, as above, and H is here a function of the estimates rather than of observed values as defined in (9). It may be noted that these results are analogous to those found in direct estimation of an observation y_j. There we have

$$(n-p-1)V(y_j-Y_j) = v_{jj}H$$

and

$$(n-p-1)\ \mathrm{Cov}\ (y_j-Y_j,y_{j'}-Y_{j'}) = v_{jj'}H$$

where the x_h are now observed quantities, the Y_j are regression estimates, and the y_j are new observations, not used in determining the regression.

The exact determination of sampling variation is not much more complicated. We may find simultaneous fiducial limits for the unknown quantities x_i in the following way. The ratio

$$\frac{n-2p}{p}\ \frac{\sum_{j}\sum_{k} v^{jk}\ (y_j - \sum_{i} b_{ij}x_i)\ (y_k - \sum_{i} b_{ik}x_i)}{H} \qquad (21)$$

is distributed as F with p and n-2p degrees of freedom. By substituting various sets of values of the x_i in the formula we can determine for which sets the associated value of F is non-significant, and hence which sets are concordant with the data. The range of concordant sets of the x_i defines a fiducial region for the values.

Now since we may write

$$y_j = \sum_{i} b_{ij}X_i \ ,$$

the y_j being observations and the X_i estimates, we have

$$\sum_j \sum_k v^{jk} \left(y_j - \sum_i b_{ij} x_i\right)\left(y_k - \sum_i b_{ik} x_i\right) = \sum_h \sum_i \sum_j \sum_k (X_h - x_h)(X_i - x_i)\, v^{jk}\, b_{hj} b_{ik}$$

$$= \sum_h \sum_i (X_h - x_h)(X_i - x_i)\, q_{hi} \qquad (22)$$

where

$$q_{hi} = \sum_j \sum_k v^{jk} b_{hj} b_{ik}$$

Since q_{hi} is a typical element of the matrix

$$Q = BV^{-1}B',$$

(22) may be written

$$(X-x)BV^{-1}B'(X'-x')$$

Hence the simultaneous fiducial limits for the values x_i are given by the solution (if real) of

$$F = \frac{n-2p}{p} \cdot \frac{\sum\sum (X_h - x_h)(X_i - x_i) q_{hi}}{H} \qquad (23)$$

If limits for a single value x_h are required, we have

$$V(X_h) = Hq^{hh}/(n-p-1) \quad . \qquad (24)$$

Now since

$$Q = BV^{-1}B',$$

$$Q^{-1} = B'^{-1}VB^{-1}$$

so that

$$q^{hi} = \sum_j \sum_k v_{jk} b^{jh} b^{ki}$$

Hence, with 1 and n-p-1 degrees of freedom,

$$F = \frac{(n-p-1)(X_h - x_h)^2}{H \sum_j \sum_k v_{jk} b^{jh} b^{kh}} \qquad (25)$$

Note that the variance estimate given by (24) differs from the approximate estimate

given above in (19) by the replacement of calculated quantities X_i by unknowns x_i. In practice, since the x_i are unknown, the approximate variance estimate based on the X_i would need to be used to give fiducial limits for a single x_h.

(b) $p < q$

In this case we have more equations than unknowns. We have a choice here either of omitting $q - p$ of the equations (provided we can decide from prior considerations which are least useful), or of using the additional information given by the equations to test the consistency of the relationships involving the different dependent variables. This latter aspect is the one that we shall examine.

If an observation of a set y_j $(j = 1, 2, \ldots q)$ of dependent variables is to be used to estimate a set x_i $(i = 1, 2, \ldots p)$, we may so determine the estimate that it has minimum (estimated) variance. Now since the estimated covariance of y_j and y_k is proportional to v_{jk}, the quantity to be minimized, with respect to the x_i, is

$$\sum_j \sum_k v^{jk} \left(y_j - \sum_i b_{ij} x_i\right) \left(y_k - \sum_i b_{ik} x_i\right)$$

If we put, as in (a)

$$Q = BV^{-1}B',$$

so that

$$q_{hi} = \sum_j \sum_k v^{jk} b_{hj} b_{ik},$$

and also put

$$P = BV^{-1}\underset{\sim}{y} \tag{26}$$

so that

$$p_i = \sum_j \sum_k v^{jk} b_{ij} y_k$$

we find for the normal equations

$$Q \underset{\sim}{X} = P \quad , \tag{27}$$

i.e.,

$$\sum_h q_{hi} X_h = p_i \quad ,$$

so that

$$\underset{\sim}{X} = Q^{-1} P \tag{28}$$

or

$$X_h = \sum_i q^{hi} p_i \quad .$$

These results are similar to those found for the case p = q, except that here the matrix B does not possess an inverse, so that the estimates need to be expressed in terms of the matrices P and Q.

As in the case p = q, the estimated covariance matrix of the X_i is

$$\frac{H}{n-p-1} Q^{-1}$$

Now we may test the consistency of the q equations by means of the departures of the observed y_j values from the estimates provided by inserting the X_i in the equations. The criterion is

$$\frac{(n-p-q)}{q-p} \frac{\sum\sum v^{jk} (y_j - \sum_i b_{ij} X_i)(y_k - \sum_i b_{ik} X_i)}{H}$$

which is distributed as F with q-p and n-p-q degrees of freedom.

This may be otherwise written

$$F = \frac{n-p-q}{(q-p)H} \left\{ \sum\sum v^{jk} y_j y_k - \sum\sum q_{hi} X_h X_i \right\}$$

$$= \frac{n-p-q}{(q-p)H} \left\{ \sum\sum v^{jk} y_j y_k - \sum\sum q^{hi} p_h p_i \right\} \tag{29}$$

If the value of F is not significant, there is no evidence for regarding the equations as inconsistent, and fiducial limits may be determined for the x_i. For these

we have

$$F = \frac{(n-p-q)}{pH} \sum\sum_{h\ i} q_{hi}(X_h - x_h)(X_i - x_i) \qquad (30)$$

with p and n-p-q degrees of freedom. This may be written in the alternative form

$$F = \frac{n-p-q}{pH} \sum\sum_{h\ h'} q^{hh'}(p_h - \sum_i x_i q_{hi})(p_{h'} - \sum_i x_i q_{h'i}) \qquad (30')$$

By means of this criterion, the concordance of any set of x_i with the data may be established.

In the particular case when p = 1, the solution of the equations gives the discriminant function for assigning a value of x_1 on the basis of observations of the q variables $y_1, y_2 \cdots y_q$.

The discriminant function is

$$\begin{aligned} X_1 &= p_1/q_{11} \\ &= \sum\sum_{j\ k} v^{jk} b_{1j} y_k \Big/ \sum\sum_{j\ k} v^{jk} b_{1j} b_{1k} \end{aligned} \qquad (31)$$

To test the consistency of any set of observations y_j, the criterion is

$$\frac{n-q-1}{q-1} \frac{\sum\sum_{j\ k} v^{jk} y_j y_k - p_1^2/q_{11}}{(1 + \frac{1}{n} + \frac{x_1^2}{t_{11}})} \qquad (32)$$

with q-1 and n-q-1 degrees of freedom.

It should be remarked here that this is a test, not of the discriminant function, which has been established from previous data, but of the consistency of the present set of observations. A significant result may indicate either that the values of y_j are not consistent among themselves, or that the discriminant function determined from previous data does not apply to the present observations.

(c) $p > q$

In this case we have fewer equations than unknowns, so that estimates of the

unknown x_i cannot be determined. The most that can be done is to find a relationship among p-q+1 of the estimates X_i. In many cases such a relationship may be all that is required, as is indicated in the example given below.

Suppose that we wish to eliminate $X_1, X_2, \ldots, X_{q-1}$, and to determine the relationship among $X_q, X_{q+1}, \ldots, X_p$. The determinant of the first q-1 rows of B and the q-1 columns resulting from omitting column j will be denoted by $(-1)^{j-1}B_j$. Then it is readily shown that the required relationship is

$$\sum_{j=1}^{q} B_j \sum_{i=q}^{p} b_{ij} X_i = \sum_{j=1}^{q} B_j y_j \qquad (33)$$

The fiducial limits for the corresponding relationship among the x_i can only approximately be determined.

The fact that $p > q$ does not, however, prevent simultaneous fiducial limits for the x_i from being found. The criterion, distributed as F with q and n-p-q degrees of freedom, from which simultaneous fiducial limits may be derived, is

$$\frac{n-p-q}{qH} \sum_j \sum_k v^{jk} \left(y_j - \sum_i b_{ij} x_i\right)\left(y_k - \sum_i b_{ik} x_i\right)$$

$$= \frac{n-p-q}{qH} \sum_h \sum_i q_{hi} (X_h - x_h)(X_i - x_i) \qquad (34)$$

where q_{hi} is the typical element of the matrix Q defined above. Here Q, though it is a p x p matrix, is of rank q.

6. Discussion of the chemical example

The original data of the experiment discussed in Section 2 are given by Fisher, Hansen and Norton (1955) in their Table I, so are not reproduced here.

Fisher et al fitted quadratic regression equations to their data, but as we found that the quadratic terms were significant only at the 5 percent level for optical densities at 560 mμ (y_2), we have ignored these terms and fitted only linear

regressions. The analyses of variance and covariance of y_1 and y_2 are shown in Table 1, and the B, T and V matrices and their inverses in Table 2.

Thus we see from Table 2 that the direct regression equations are

$$\begin{aligned} Y_1 &= 1.2166\, x_1 + 2.6240\, x_2 \\ Y_2 &= 1.3465\, x_1 + 4.7276\, x_2 \end{aligned} \tag{35}$$

and the inverse equations are

$$\begin{aligned} X_1 &= 2.1311\, y_1 - 1.1829\, y_2 \\ X_2 &= -0.6070\, y_1 + 0.5484\, y_2 \end{aligned} \tag{36}$$

in agreement with the results of Fisher *et al.*

The direct equations are less useful than the inverse ones. Since in this example the numbers of dependent and independent variables are equal, no test for consistency is possible, but we can derive fiducial limits for the values of x_1 and x_2 corresponding to observed values y_1 and y_2.

Since the inverse regression coefficients, as well as the direct coefficients, are likely to be well determined, we may calculate their approximate standard errors. Table 3 gives the matrices $B^{-1}T^{-1}B'^{-1}$ and $B'^{-1}VB^{-1}$ required in these calculations. Then, for example, the variance of b^{21} is obtained using the second diagonal term of $B^{-1}T^{-1}B'^{-1}$ and the first diagonal term of $B'^{-1}VB^{-1}$:

$$9.183 \times 10^{-6} \times 8699/26 = 0.003072,$$

so that the standard error of b^{21} is 0.055. The standard errors of the coefficients may be set out as follows:

$$\begin{bmatrix} 0.091 & 0.034 \\ 0.055 & 0.021 \end{bmatrix}$$

For general purposes, of course, the covariances as well as the variances of the regression coefficients will be of interest.

In determining the approximate variance of an estimate X_i, since the regression is through the origin rather than the point of means, the actual values of the y_j rather than departures from means are used, and the term $\frac{1}{n}$ is omitted from the variance estimates, in equation (19).

Thus, approximately,

$$V(X_1) = \frac{10^{-6}x8699}{26} (1 + 24.99\ y_1^2 - 30.06\ y_1y_2 + 9.18\ y_2^2)$$

$$= \frac{10^{-6}x8699}{26} (1 + 4.396\ X_1^2 - 2.637\ X_1X_2 + 4.396\ X_2^2)$$

with similar results for Cov (X_1, X_2) and $V(X_2)$.

7. An example of inverse estimation where $p > q$

A study of the pulping properties of eucalypt woods is reported by Cohen and Mackney (1951). The object of the studies was to determine a treatment which would produce pulp of the required lignin content, from wood with certain characteristics. The percentage of the wood material soluble in hot water (hot-water solubles, x_1) was determined for each wood sample, which was then divided into four parts, each being pulped with varying amounts of active alkali (x_2 percent of the wood weight). The same levels of active alkali were repeated for each sample, so that the two independent variables were uncorrelated. The lignin content of the resulting pulp was measured in terms of a "permanganate number", and its logarithm to base 10(y) taken as the dependent variable. The data are shown in Table 4.

This example does not illustrate the use of simultaneous equations, but it does show how inverse estimation is possible when there are more independent than dependent variables. The appropriate regression is that of y on x_1 and x_2; however, what is required from the data is an estimate of the relationship of x_2 to x_1, corresponding to a fixed value of y; in other words, the alkali requirement x_2 which

will result on the average in a given lignin content Y, when the hot-water solubles figure x_1 is known.

The relevent sums of squares and products are shown in Table 5. The regression equation is found to be

$$Y = 2.123 + 0.0301x_1 - 0.0596x_2. \qquad (37)$$

The analysis of variance in Table 6 shows this regression to be highly significant, and the residual variance to be 0.005 804. Since the 1 percent point of F with 1 and 57 degrees of freedom is 7.102, the 99 percent fiducial boundary for the regression relationship is

$$(Y - 2.123 - 0.0301x_1 + 0.0596x_2)^2$$
$$= 7.102 \times 0.005\,804 \left(\frac{1}{60} + \frac{(x_1 - 7.486)^2}{516.315} + \frac{(x-18)^2}{300}\right) \qquad (38)$$

The lignin content required for the pulp corresponds to a "permanganate number" of 15 (i.e., Y = 1.176); this value inserted in the equation gives the relationship

$$X_2 = 15.89 + 0.504x_1 \ ,$$

so that, once the hot-water solubles percentage is given, the requirement of active alkali can be estimated. The fiducial boundary for the relationship is given by substituting y = 1.176 in equation (38).

8. Proportional regressions

In certain cases it is of interest to fit equations in which the coefficients are proportional. For instance, in studying various properties of coals, such as their carbon content, sulphur content and calorific value, it may be supposed that each is linearly related to the percentage ash content. It might be expected that, if the ash were simply the result of admixed impurities, its effect would be a simple percentage reduction, the same for each of the properties. Thus, if x were the ash

content, and the regression of the jth property y_j on x were

$$Y_j = b_{oj} + b_j x \quad , \tag{39}$$

we should expect b_j to be negative, and b_j/b_{oj} to be in the neighbourhood of -1/100. In general, if the theoretical value of the ratio were $-1/\xi$, we could fit the restricted regression equations

$$Y_j = b_j'(x - \xi) \tag{40}$$

the value of ξ being the same for each line.

We should then be interested in testing, firstly, the validity of the assumption of a constant value ξ for the different dependent variables, and secondly, the acceptability of various values of ξ.

We shall here consider only the case of one independent variable, which seems to be of most practical interest; the extension to more than one independent variable introduces no new principle. The additional complication in fitting proportional equations arises from the fact that there is a constant common to all the equations.

Since the equations of estimation of ξ are not linear, the method of least squares does not lead to exact significance tests and fiducial limits. Instead, the following method is adopted. The null hypothesis on which the test is based is that the regressions are proportional, with constant of proportionality ξ. The constant is unspecified, so that the test criteria are functions of ξ. If for any particular value of ξ the test criteria are significant, the null hypothesis, and the corresponding value of ξ, are rejected at the level of significance adopted. We are thus able to set fiducial limits on ξ.

We shall denote the sum of squares of x by t, and the sum of products of y_j with x by p_j, and shall adopt the following notation for the restricted regressions with constant ξ:

$$p_j' = Sy_j(x - \xi)$$

$$t' = S(x - \xi)^2$$

$$b_j' = p_j'/t'$$

To test the validity of the hypothesis, consider the unrestricted regressions, which may be written

$$Y_j = \bar{y}_j + b_j (x-\bar{x}) \quad .$$

When $x = \xi$, Y_j should differ from zero only by errors of random sampling. Hence the q quantities

$$z_j = \bar{y}_j + b_j (\xi - \bar{x}) \tag{41}$$

have a joint normal distribution centered on zero. The analysis of these quantities provides tests of the hypothesis.

The covariance of z_j and z_k, estimated with n-2 degrees of freedom, is

$$\frac{t'}{nt} \frac{v_{jk}}{(n-2)}$$

so that the sum of squares of these quantities may be taken as

$$\frac{nt}{t'} \sum_j \sum_k v^{jk} z_j z_k \tag{42}$$

This is distributed as the ratio of two independent sums of squares with q and n-q-1 degrees of freedom, so may be tested directly by means of the F distribution. We have

$$F = \frac{(n-q-1)nt}{qt'} \sum_j \sum_k v^{jk} z_j z_k \tag{43}$$

This provides an overall test of the assumption of proportionality and of the specified value of the constant. This may be partitioned to give tests separately of the two aspects.

Now the quantities b_j', or the equivalent p_j', can be shown to represent the

variation among the restricted regressions. Also, since

$$p_j' = p_j - n\bar{y}_j\,(\xi-\bar{x}) \tag{44}$$

we see that the expected value of the correlation of z_j and p_j' between lines is zero. This is an expression of the fact that the p_j' account for all, and the z_j for none, of the variation between lines.

Hence the sample regression of the z_j on the p_j' provides a test criterion for the hypothetical value of ξ.

The sum of squares for regression is

$$\frac{nt\left[\sum\sum_{j\ k} v^{jk} z_j p'_k\right]^2}{t'\sum\sum_{j\ k} v^{jk} p'_j p'_k} \tag{45}$$

This is distributed as the ratio of two independent sums of squares with 1 and n-q-1 degrees of freedom, and thus may be tested by the F distribution.

$$F = \frac{(n-q-1)nt\left[\sum\sum_{j\ k} v^{jk} z_j p'_k\right]^2}{t'\sum\sum_{j\ k} v^{jk} p'_j p'_k} \tag{46}$$

The sum of squares, with q-1 and n-q-1 degrees of freedom, for departures of the z_j from regression on the p_j', which is given by the difference between (42) and (45), is available for testing departure from proportionality.

It is convenient to express these sums of squares and products in a form which shows explicitly their dependence on ξ. If we write

$$J = n\sum\sum_{j\ k} v^{jk}\bar{y}_j\bar{y}_k$$

$$K = \frac{n}{t}\sum\sum_{j\ k} v^{jk}\bar{y}_j p_k$$

$$L = \frac{n}{t^2}\sum\sum_{j\ k} v^{jk} p_j p_k$$

then the total sum of squares (42) of the z_j is

$$\frac{t}{t'}\left[J + 2(\xi-\bar{x})K + (\xi-\bar{x})^2 L\right] \tag{47}$$

while the sum of squares (45) for regression of z_j on p_j' is

$$\frac{nt\left[(\xi-\bar{x})^2 K - (\xi-\bar{x})(\frac{t}{n}L-J) - \frac{t}{n}K\right]^2}{t'\left[n(\xi-\bar{x})^2 J - 2(\xi-\bar{x})tK + \frac{t^2}{n}L\right]} \tag{48}$$

The sum of squares for departure from proportionality is found, by subtraction, to be

$$\frac{t\,t'\left[JL-K^2\right]}{n\left[n(\xi-\bar{x})^2 J - 2(\xi-\bar{x})tK + \frac{t^2}{n}L\right]} \tag{49}$$

The full analysis may be set up in the form of an analysis of variance, as follows:

	Degrees of Freedom	Sum of Squares
Constant of proportionality	1	$\dfrac{nt\left[(\xi-\bar{x})^2 K-(\xi-\bar{x})(\frac{t}{n}L-J)-\frac{t}{n}K\right]^2}{t'\left[n(\xi-\bar{x})^2 J-2(\xi-\bar{x})tK+\frac{t^2}{n}L\right]}$
Departure from proportionality	q-1	$\dfrac{t\,t'\left[JL-K^2\right]}{n\left[n(\xi-\bar{x})^2 J-2(\xi-\bar{x})tK+\frac{t^2}{n}L\right]}$
Error	n-q-1	1
Total	n-1	$1 + \frac{nt}{t'}\sum_j\sum_k v^{jk} z_j z_k$
		$= \left\| v_{jk} + \frac{nt}{t'} z_j z_k \right\| \Big/ \left\| v_{jk} \right\|$

Provided the departure from proportionality is not significant, the analysis enables us to test a specified value of ξ, or, more generally, to set fiducial

limits. An estimate of the constant of proportionality is one of the two values of ξ which make the sum of squares for constant of proportionality vanish. Now one of these values minimizes the sum of squares for departure from proportionality, while the other maximizes it. The former value is the appropriate estimate. It will be noted that, considered as a function of ξ, the sum of squares for departure from proportionality is inversely proportional to

$$\frac{1}{t'}\left[n(\xi-\bar{x})^2 J - 2(\xi-\bar{x})tK + \frac{t^2}{n}L\right] \tag{50}$$

which is the sum of squares for differences among the p_j', or for difference of regressions.

Thus the optimum estimate of ξ is that which minimizes departure from proportionality and maximizes difference of (proportional) regressions.

If the X is the estimate of ξ, the equation for X is

$$(X-\bar{x})^2 K - (X-\bar{x})\left(\frac{t}{n}L-J\right) - \frac{t}{n}K = 0$$

giving

$$X = \bar{x} + \frac{\frac{t}{n}L - J \pm \sqrt{\left[\left(\frac{t}{n}L+J\right)^2 - \frac{4t}{n}(JL-K^2)\right]}}{2K} \tag{51}$$

The two roots lie on opposite sides of $\bar{x}$.

In other contexts, an analysis similar to this one will provide a test for the constancy of a set of ratios, and fiducial limits for their common expected value.

Table 1

Analyses of Variance and Covariance of Optical Density Measurements at 470 mμ (y_1) and at 560 mμ (y_2) (Fisher, Hansen and Norton's data)

	Degrees of Freedom	Sums of Squares and Products		
		y_1^2	y_1y_2	y_2^2
Regression on x_1, x_2	2	2.570 253	4.207 267	6.995 805
Residual	26	0.003 167	0.002 996	0.006 733
Total	28	2.573 420	4.210 263	7.002 538

Table 2

Matrices of Sums of Squares and Products, and of Regression Coefficients

$$T = \begin{bmatrix} 0.2500 & 0.0750 \\ 0.0750 & 0.2500 \end{bmatrix} \qquad 10^6V = \begin{bmatrix} 3167 & 2996 \\ 2996 & 6733 \end{bmatrix} \qquad B = \begin{bmatrix} 1.2166 & 1.3465 \\ 2.6240 & 4.7276 \end{bmatrix}$$

$$T^{-1} = \begin{bmatrix} 4.3956 & -1.3187 \\ -1.3187 & 4.3956 \end{bmatrix} \qquad V^{-1} = \begin{bmatrix} 545.3 & -242.6 \\ -242.6 & 256.5 \end{bmatrix} \qquad B^{-1} = \begin{bmatrix} 2.1311 & -0.6070 \\ -1.1829 & 0.5484 \end{bmatrix}$$

Table 3

Matrix Products Required in Estimating Variances

$$M^{-1} = B^{-1}T^{-1}B'^{-1} = \begin{bmatrix} 24.995 & -15.032 \\ -15.032 & 9.183 \end{bmatrix} \qquad 10^6(Q^{-1} = B'^{-1}VB^{-1}) = \begin{bmatrix} 8699 & -2812 \\ -2812 & 1197 \end{bmatrix}$$

Table 4

Data from a Study of Pulping Properties of Eucalypt Woods

x_1 = percentage hot-water solubles

x_2 = percentage active alkali used in pulping

y = log permanganate number

x_1	x_2	y	x_1	x_2	y	x_1	x_2	y
5.97	15	1.425	6.79	15	1.498	13.19	15	1.734
	17	1.250		17	1.330		17	1.535
	19	1.170		19	1.233		19	1.326
	21	1.124		21	1.161		21	1.201
8.00	15	1.641	9.20	15	1.442	9.52	15	1.500
	17	1.418		17	1.255		17	1.281
	19	1.230		19	1.146		19	1.152
	21	1.164		21	1.093		21	1.104
8.51	15	1.655	10.00	15	1.507	9.46	15	1.610
	17	1.384		17	1.332		17	1.425
	19	1.334		19	1.220		19	1.283
	21	1.164		21	1.199		21	1.204
4.51	15	1.486	10.94	15	1.667	3.17	15	1.204
	17	1.272		17	1.458		17	1.130
	19	1.185		19	1.258		19	1.083
	21	1.124		21	1.173		21	1.004
3.15	15	1.250	6.35	15	1.391	3.53	15	1.236
	17	1.146		17	1.207		17	1.149
	19	1.086		19	1.100		19	1.061
	21	1.033		21	1.079		21	1.025
Total	15	22.246						
	17	19.572			Means	7.486	18	1.2756
	19	17.867						
	21	16.852						
Grand Total		76.537						

Table 5

Calculation of Regression Coefficients from Values in Table 4

	Sum of squares	Sum of products with y	Regression coefficient	Regression sum of squares
x_1	516.315	15.5292	+ 0.030077 ± 0.0034	0.4671
x_2	300	-17.887	- 0.059623 ± 0.0044	1.0665
				1.5336

Table 6

Analysis of Variance

	d.f.	Sum of squares	Mean square
Regression	2	1.5336	0.7668**
Residual	57	0.3308	0.005 804
Total	59	1.8644	

References

Box, G.E.P. and Hunter, J.S. (1954). A confidence region for the solution of a set of simultaneous equations with an application to experimental design. Biometrika 41, 190-199.

Cohen, W.E. and Mackney, A.W. (1951). Influence of wood extractives on soda and sulphate pulping. Aust. Pulp Paper Ind. Tech. Assoc. Proc. 5, 315-335.

Fisher, Hans, Hansen, R.G. and Norton, H.W. (1955). Quantitative determination of glucose and galactose. Anal. Chem. 27, 857-859.

van der Waerden, B.L. (1931). Moderne Algebra. Berlin. Springer.