AN APPLICATION OF LEAST SQUARES TO FAMILY DIET SURVEYS[1]

By M. H. Quenouille

The application of the method of least squares is discussed for the case in which a large number of unknowns have to be estimated; in particular, the application to family diet surveys is considered. Methods are indicated for shortening the analysis and investigating the importance of different variables by a step-by-step procedure. Suggestions are made on the planning of surveys in the light of the methods investigated here.

SUMMARY

In any well-planned diet survey, certain basic information is collected from each family observed. Briefly this includes (a) the total food consumption of the family during a given period, derived by weighing foods in stock at the beginning and end of the period and foods purchased during the period, (b) the age and sex distribution of the family and the occupations of all wage earners, and (c) the number of meals taken out by members of the family and the number of meals consumed within the household by visitors. (It is rarely possible to determine the exact outside consumption and consumption by visitors.)

From these data calculations are usually made of the nutritive value of the home diet in terms of energy value and of certain constituents, and of the requirements of the family for energy and the same constituents in terms of some accepted scale. Comparisons can then be made between population groups, usually with certain additional devices such as stratification by food expenditure per head of each family. Under the conditions prevailing up to 1939, the income per head of a family was the main determinant of food expenditure, and that in turn of quality of diet. Such a stratification therefore gave information valuable from the economic, sociological, and nutritional aspects. However, it is incapable of answering many subsidiary questions, such as the effects on food consumption of occupation, and family size and composition, except in so far as these tend to coincide with the economic classification.

When I was asked to undertake a study of the data from the Carnegie United Kingdom Dietary Survey (1937–39), the suggestion was made that I should explore to what extent additional information could be obtained about the effect of occupation and family composition on consumption, and whether or not any significant information could be derived about the consumption of individuals. Previously, only one

[1] My thanks are due to Dr. I. Leitch of the Commonwealth Bureau of Animal Nutrition for assistance in the preparation of this paper.

attempt had been made to assess individual consumption from family surveys, that of Clements [1], in which he gauged the individual consumptions by picking out and comparing families of similar constitution differing in one number. This was obviously unsatisfactory since it utilised only a small fraction of the available information and was subject to very large errors.

The Carnegie Survey was not planned with this type of refined analysis in view, but it has been found possible to obtain a great deal of significant information from the data. It would have been easier to analyse, and even more could have been extracted from it, had the survey been suitably planned.

Since the dietary survey records food purchases for the family as a unit, the results as referred to individuals will not necessarily give exactly the same answer as would individual dietary surveys. What is studied is the effect of family constitution and all the other variables on food purchasing habits. That is to say, the result represents a general pattern and not the findings for actual individuals.

At the same time, the form of analysis employed on the data of this survey indicates a method of dealing with least squares equations when a large number of variables are involved. Thus, in the following sections, a method is suggested of carrying out an analysis so that the major causes of variation can be removed and examined in a step-by-step procedure, and, in particular, the major causes of variation in family diet surveys are examined and examples are given.

1. INITIAL ANALYSIS

In general, we shall have a series of figures for the total consumption per family over a period of time, which figures it will be necessary to split into individual components. We can assign constants to each member of the family representing his or her daily consumption. Thus a_1 might be used to denote the consumption of a child, age 0–1, not breast-fed, and similarly constants may be used for other members of the family, say, as follows:

a_2 breast-fed child, age 0–1,
b child, age 1–3,
c child, age 4–6,
d child, age 7–9,
e child, age 10–12,
f_1 male adolescent, age 13–15,
f_2 female adolescent, age 13–15,
g_1 male adolescent, age 16–20,
g_2 female adolescent, age 16–20,

m combined consumption of first male and female adults,
n consumption of each extra male adult,
p consumption of each extra female adult.

The age grouping is in accordance with the League of Nations scale of requirements, but any alternative grouping may be used. It must be remembered, however, that the use of additional constants requires a large amount of extra work and that this work would be in vain if the amount of data did not justify an extensive calculation. For this reason the grouping of the adults according to age will seldom be worthwhile, although it is not difficult to test this after a preliminary analysis. It should be noted that individual constants to represent adult male and female consumption separately have not been used. This is because the majority of families contain just one male adult and female adult member, and a deviation from this condition frequently represents a state of economic stress in the family. To base individual male and female figures on such cases would lead to spurious conclusions, although in this case the extra adults might be grouped according to age or some other classification to differentiate between two families living together and one family with more than two adult members. It is more important to classify the adult wage earners of the family according to the type of work that is being carried out. Thus instead of m, four additional constants, m_1, m_2, m_3, and m_4, might be used for the respective cases in which the male member is a heavy worker, medium worker, light worker, or unemployed. A similar classification might be used for n, but the relatively small number of cases in which families have more than one male adult member does not generally warrant the extra work involved.[2]

Using this system, for example, $m_2 + a_1 + 2c + d$ would represent one week's consumption by a family of one male medium worker, one adult female, and children aged 0–1, 4–6, 4–6, and 7–9. In the method of analysis proposed below, it will be presupposed that all families are observed for the same length of time. This is usually true, and a week is usually taken as the survey period, which implies payment of wages weekly and the usual absence of shopping facilities on Sunday.

In any analysis it will be necessary to take into account causes of variation other than family constitution. Five main causes other than family constitution exist:

Location. The manner in which the location of the family is taken into account will depend upon the importance that is placed upon district-to-district comparisons. Obviously if fine points are to be investigated, then a large number of observations must be taken and each district analysed

[2] Under present conditions, a large number of families are sharing accommodations and food, and this would no longer be true.

separately, but in most cases the object will be to find the difference in consumption per individual. For this purpose it will suffice to add constants representing the relative individual consumption for each district or type of district. For example, if we group the districts into rural, light industrial, and heavy industrial, constants u, v, and w, $u + v + w = 0$, might be used to represent the relative individual consumptions and thus $m_2 + a_1 + 2c + d + 6v$ would represent the above family in a light industrial district. It will be seen that this representation fails in that it assigns only one constant to each district comparison irrespective of age or occupation. To be most effective these constants should take into account differences in age and occupation; but, provided that the families used in each analysis are of roughly the same size, the error introduced will be insignificant if we remember that adults will account for more, and children less, than is indicated by the values that are deduced.

Food expenditure or income group. Allowance can be made for economic differences in the same manner as for differences in location, but the extreme differences in this case are so large that it would seem advisable, if the data permit, to analyse them in more than one portion, although this is a difficulty that can be overcome by the use of further constants.

Size of family. The variation of individual intake with size of family might be expected, and this might be studied by the introduction of further constants (see below), but the increasing variation in family consumption with size of family necessitates separate analyses for increasing size of family, prior to any over-all analysis, to determine the relative accuracy of results obtained from families of different sizes. It would seem likely that in most cases the variation would increase with size of family, leaving the coefficient of variation constant; and in some recent unpublished work on calorie intake this was in fact true, the coefficient of variation remaining at roughly 15% throughout when all definable causes had been eliminated. Thus the results from families of different sizes must be weighted according to some preliminary analysis or some predetermined scale before a final analysis is undertaken.

Seasonal variation. The seasonal variation in consumption is relevant to any analysis, particularly of fruit and vegetables, but does not cause appreciable trouble in the analysis, although this effect will often vary with size of family. Usually in the planning of the survey this effect will be made orthogonal to most other effects.

Visitors and meals out. The manner in which visitors and meals out are dealt with must depend upon what is intended to be measured by the survey. If it is the individual consumption at home that is to be measured then no adjustment for meals out is necessary, and such an adjustment would seem generally undesirable, especially when rationing

is in force. An adjustment for visitors can be made, but unfortunately the case of the casual visitor (when further food is introduced into the house) cannot be classified separately from the case of the "unexpected caller" (when little or no extra food is introduced) or of the neighbor who returns the compliment. Some distinction is preferable, although, since any constant will take on an average value for the effects of different types of visitors and since the proportion of visitors is in any case usually small, such a distinction is not important.

2. THE BASIC METHOD OF ANALYSIS

When the "model" for analysis has been specified, then it is next necessary to determine the form and method of analysis. The analysis can be carried out conveniently by the method of least squares, which minimizes $\sum\omega_i(x_i - X_i)^2$, where x_i is the mean daily consumption of the ith family, X_i is the expected daily consumption of the ith family in terms of unknown constants, and ω_i is the weight attached to the observation depending on the size of family and possibly the length of observation.

This well-known method gives rise to as many simultaneous linear equations as there are unknown constants, and, if the accuracy of the estimated constants is to be found, a matrix of the same order must be inverted. It is for this reason that it is necessary to reduce the number of constants to a minimum. The calculation, however, can be reduced since the unknowns can be regarded as several sets of constants. For example, either m_1, m_2, m_3, or m_4 occurs in every family and, similarly, either u, v, or w occurs in every income group classification. With two such sets it has been shown [2] that an analysis of variance can be carried out on one set and adjusted by covariance on a dummy variate representing the other. This can similarly be carried out with three or more sets, although it requires the use of covariances on covariances which will be discussed later. However this suggests methods of planning a survey so that the subsequent statistical analysis can be greatly shortened. The simplest method would be to ensure that district, income group, and type of worker comparisons are all orthogonal by taking the same proportions of each type of worker in each district and income group and by taking the same proportions of each income group from each district, or, if each income group is to be analysed separately, by taking the type of worker and district orthogonal to each other. Unfortunately this is not always practicable because the interdependence of the factors involved makes the collection of such samples difficult in many cases, unreal, and inefficient for the purpose of subsequent over-all comparisons. The alternative, which makes a compromise between the statistical analysis and the efficiency of individual compari-

sons on the one hand, and the ease of sampling and the efficiency of over-all comparisons on the other, seems to demand the use of "partially-orthogonal" samples. An approach to such samples can be made by the covariance method mentioned above. When there are two criteria of classification then an analysis of one criterion with covariance on dummy variates representing the other will involve a covariance with one less variable than is involved in the second criterion. However this covariance can be partitioned in different ways into orthogonal components, and if the majority of these components are taken to be orthogonal to the first criterion, then these can be taken out in the initial analysis and the covariance carried out on the remaining components. For example, suppose that the second criterion splits the data into three groups, u, v, w, say, and that the number of observations in the subgroups are

TABLE I

SECOND CRITERION	GROUPS OF FIRST CRITERION					
	1	2	3	4	5	...
u	n_1	n_2	n_3	n_4	n_5	...
v	m_1	m_2	m_3	m_4	m_5	...
w	$m_1p - n_1$	$m_2p - n_2$	$m_3p - n_3$	$m_4p - n_4$	$m_5p - n_5$	...

those given in Table I. It is seen that an initial analysis of the first criterion and of $u + w$ versus v can be carried out and adjusted by a single covariance using the dummy variate ξ taking values 1 for u, 0 for v, and -1 for w.

When three or more criteria of classification are used then the same principle can again be employed, although its application is more involved and depends upon whether or not the several criteria are to be taken as mutually orthogonal. For example, if a third criterion with the grouping x, y, z, is employed with numbers in the subgroups as in Table I, then $u + w$ and v must be orthogonal to $x + z$ and y for a second dummy variate to be employed.

A final simplification in the analysis that might be employed is the assumption that the consumption by the children behaves in a "regular" manner, i.e., its differentials exist everywhere and it can be adequately represented by a polynomial. This is not always possible, but it provides a useful simplification, especially when, as is common, the estimates in successive age groups tend to be negatively correlated. In any case, some care must be exercised in using this method to ensure that the representation is adequate, and if the sample is large there would seem little point in its use. In the following example this method has been used

since the purpose is to demonstrate the analysis, and on as small a sample as sixty observations the estimates of the constants would be subject to very large error.

Example. To demonstrate the analysis, sixty observations, C, of calorie consumption of families with five children and spending about the same amounts on food per head were taken from a pre-war survey. Originally

TABLE II

Type of Worker	Type of District								
	Heavy Industrial $\xi = -1$			Light Industrial $\xi = 0$			Rural $\xi = 1$		
	l	q	C	l	q	C	l	q	C
Heavy	− 3	15	12685	− 5	9	15712	3	19	13668
	− 2	6	14513	0	6	15056	− 7	15	16639
	− 7	15	11478	3	19	15314	− 3	11	14758
	− 1	9	14867	− 3	15	12844	− 5	9	15451
	−13	45	53543	− 5	49	58296	−12	54	60516
Medium	− 6	14	11868	1	11	16466	2	10	15447
	− 1	11	12194	− 1	7	14291	− 1	7	15970
	1	9	15659	− 3	15	14527	− 2	4	14158
	− 3	15	14016	− 3	15	12480	− 6	10	16034
	1	9	13343	0	10	17569	− 1	7	17737
	− 3	7	11391	− 7	15	12757	− 7	15	15450
	− 5	9	13410	2	14	18002	− 6	10	15444
	5	15	17092	− 4	10	15144	− 1	11	16473
	2	14	15118	− 3	7	13275	− 1	6	19622
				− 6	10	14100	− 5	9	20552
							− 3	5	19830
	− 9	103	124091	−24	114	148611	−31	94	186717
Light	− 6	14	11777	− 1	7	15416			
	− 3	7	14200	− 1	11	20989			
	0	8	18300						
	1	11	13893						
	− 8	40	58170	− 2	18	36405			
Unemployed	− 8	16	12102	− 6	14	15246	− 4	18	11809
	− 1	15	11340	− 3	13	13166	− 4	10	16166
	− 2	6	13425	− 2	6	17502			
	− 4	16	14057	5	15	16275			
	− 7	15	13658						
	5	15	17667						
	−17	83	82296	− 6	48	62189	− 8	28	27975
Total.....	−47	271	318100	−37	229	306131	−51	176	275208

there had been about eighty observations falling into this category, but observations on families with breast-fed children, pregnant women, or more than two adults, were rejected, and a further group of 17 observations was rejected to bring the remainder into a pattern similar to that indicated in Table I. In this form the equations to be solved involve only one constant instead of the seven, m_1–m_4, u, v, and w. Since the number of children in each family is the same, no comparison of adult consumption with consumption per child is possible, and the small number of observations precludes all but the grossest comparisons. Thus

TABLE III

PRELIMINARY ANALYSES OF COVARIANCE

	Degrees of Freedom	Sum of Squares of C	Sum of Squares of ξ	Sum of Products of ξ and l	Sum of Products of ξ and q	Sum of Products of ξ and C
Mean	1	13,483,175,245	0.6000	13.5000	−67.6000	−89,944
Workers	3	15,278,930	3.5333	− 0.7676	− 3.3333	35
$w + u - 2v$	1	2,993,784	0.3000	1.2000	0.5500	948
Residual	55	288,369,561	35.5676	−17.9333	−24.6176	46,069
	60	13,789,817,520	40.0000	−4.0000	−95.0000	−42,892

		Sum of Squares of l	Sum of Products of l and q	Sum of Products of l and C
Mean	1	303.750	−1521.000	−2023738
Workers	3	4.533	− 19.883	8035
$w + u - 2v$	1	4.800	2.200	3791
Residual	55	609.917	2.683	125079
	60	923.000	−1536.000	−1886833

		Sum of Squares of q	Sum of Products of q and C
Mean	1	7616.267	10133679
Workers	3	100.517	− 38244
$w + u - 2v$	1	1.008	1737
Residual	55	760.208	−131110
	60	8478.000	9966062

If the predicted calorie consumption is given by $a_1\xi + a_2 l + a_3 q$, then, from Table III,

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 35.667 & -17.933 & -24.617 \\ -17.933 & 609.917 & 2.683 \\ -24.617 & 2.683 & 760.208 \end{bmatrix}^{-1} \begin{bmatrix} 46069 \\ 125079 \\ -131110 \end{bmatrix}$$

$$= \begin{bmatrix} 0.0291991 & 0.0008544 & 0.0009425 \\ 0.0008544 & 0.0016646 & 0.0000218 \\ 0.0009425 & 0.0000218 & 0.0013459 \end{bmatrix} \begin{bmatrix} 46069 \\ 125079 \\ -131110 \end{bmatrix} = \begin{bmatrix} 1328.475 \\ 244.709 \\ -130.311 \end{bmatrix}.$$

The analysis of covariance is then completed as follows:

	Degrees of Freedom	Sum of Squares of C	Mean Square	Standard Deviation
Regression	3	108,894,891		
Residual (2)	52	179,474,670	3,451,436	1,858
Residual (1)	55	288,369,561		

We may now test different effects, for example,

(1) Standard error of $a_1 = [(0.0291992)\ (3,451,436)]^{1/2} = 317.5$,
Standard error of $a_2 = [(0.0016646)\ (3,451,436)]^{1/2} = 75.8$,
Standard error of $a_3 = [(0.0013459)\ (3,451,436)]^{1/2} = 68.2$.

(2) Difference in calorie intake of children, aged 16–20 and 7–9 equals $3a_2 + 9a_3 = -438.7$.
Standard error of difference equals $[(9 \cdot 0.0016646 + 54 \cdot 0.0000218 + 81 \cdot 0.0013459)\ 3,451,436]^{1/2} = 688.4$

(3) Difference between mean calorie intake of medium and unemployed workers, adjusted for differences in family constitution and district equals $15314 - 14372 - (24a_1 + 27a_2 - 173a_3)/60 = -75$.
Standard error of difference equals $[(1/30 + 1/12 + 48.06/3600)\ 3,451,436]^{1/2} = 669.8$.

the substitutions

$$a_2 = d - 3l + 9q, \qquad b = d - 2l + 4q, \qquad c = d - l + q,$$

$$e = d + l + q, \quad f_1 = f_2 = d + 2l + 4q, \quad q_1 = q_2 = d + 3l + 9q,$$

were made. Table II gives the mean daily consumption per family, together with the coefficients of l and q for each family.[3] The covariances

[3] My thanks are due to Dr. D. P. Cuthbertson of the Rowett Institute for Research in Animal Nutrition for permission to use the data in this table from the Carnegie Dietary Survey prior to the publication of the analysis of this survey.

are then analysed in Table III and in the material which follows immediately thereafter. The analysis thus follows the normal course and standard errors can be attached to the resulting estimates and their differences in the normal way. This analysis, although crude in many ways, demonstrates on a smaller scale how the method works, and it also emphasizes a difficulty that will often occur: Is the "model" used for the analysis sufficiently good, i.e., what additional factors, if any, should be introduced and how can they be introduced with a minimum of labour?

3. COVARIANCES ON COVARIANCES

The above analysis was made without undue difficulty because it was assumed that all the possible influencing factors had been taken into account and that the number of such factors was relatively small. However, in practice this is not always true and, as pointed out in the last section, we are frequently uncertain whether or not additional variables should be added when the number of variables is very large. Consequently, although the above analysis saves a great deal of time, the computation may still be very large. For example, we may have twenty prime causes of variation and another ten possible causes. By preliminary planning, ten of the prime causes may be made orthogonal, but a covariance analysis still involves the inversion of a 10 by 10 matrix with further inversions if the possible causes of variation are also to be tested. This work can again be shortened using the "covariance-on-covariance" technique. Thus, suppose that the residuals in the initial analysis of variance are represented by y and that the initial covariance on $x_1, \cdots, x_p$ estimates the relation $Y = a'x$, where $a' = [a_1, \cdots, a_p]$ and $x' = [x_1, \cdots, x_p]$. Let $s(\xi_i\eta)$ represent the column vector with elements $\sum\xi_i\eta$ and let $S(\xi_i\eta_j)$ represent the matrix with elements $\sum\xi_i\eta_j$. Also let $S^{-1}(\xi_i\eta_j) = C(\xi_i\eta_j)$; then $a = C(x_ix_j)s(x_iy)$.

If further variables $z_1, \cdots, z_q$ are now introduced, then regressions on $x_1, \cdots, x_p$ give $Z_k = S'(z_kx_i)\,C\,(x_ix_j) = l_k'x$, say, and the residual sum of squares and products for $y, z_1, \cdots, z_q$ when $x_1, \cdots, x_p$ have been eliminated are given by formulae such as

$$\begin{aligned}
\Sigma_x y^2 &= \Sigma y^2 - S'(x_iy)C(x_ix_j)S(x_iy) &&= \Sigma y^2 - a'S(x_iy),\\
\Sigma_x z_1^2 &= \Sigma z_1^2 - S'(x_iz_1)C(x_ix_j)S(x_iz_1) &&= \Sigma z_1^2 - b_1'S(x_iz_1),\\
\Sigma_x yz_1 &= \Sigma yz_1 - S'(x_iy)C(x_ix_j)S(x_iz_1) &&= \Sigma yz_1 - a'S(x_iz_1)\\
& && = \Sigma yz_1 - b_1'S(x_iy),
\end{aligned}$$

where the subscript x is used to indicate that the variables $x_1, \cdots, x_p$ have been eliminated. Thus, if a further covariance on $z_1, \cdots, z_q$ is carried out, the inverse matrix is $C_x(z_iz_j)$ and we now have

$$Y_x = S_x'(yz_i)C_x(z_iz_j)z_x = d'z_x,$$

and the residual sum of squares for y is given by

$$\Sigma_{zz}y^2 = \Sigma_x y^2 - d'S_x(yz_i).$$

If we write out the relation connecting Y with $x_1, \cdots, x_p, z_1, \cdots, z_q$, we now get

$$Y = d'z + (a' - dB)x = d'z + e'x,$$

where

$$B' = [b_1, b_2, \cdots, b_q] = S'(z_k x_i)C(x_i x_j).$$

The covariance matrices of $d_1, \cdots, d_q$ and $a_1, \cdots, a_p$ are $C_x(z_i z_j)$ and $C(x_i x_j)$, respectively, so that the covariance matrix of $e_1, \cdots, e_p$ is

$$C(x_i x_j) + B'C_x(z_i z_j)B$$

and the covariance of $d_1, \cdots, d_q$ with $e_1, \cdots, e_p$ is given by

$$-B'C_x(z_i z_j).$$

Thus the over-all covariance matrix of $e_1, \cdots, e_p, d_1, \cdots, d_q$ is

$$\begin{bmatrix} C(x_i x_j) + B'C_x(z_i z_j)B & \cdots & -B'C_x(z_i z_j) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & & \\ -C_x(z_i z_j)B & \cdots & C_x(z_i z_j) \end{bmatrix}.$$

The importance of this approach, which effectively extends a formula given by Cochran [3] for the addition of an extra variable in a regression, is that the over-all analysis involving $p + q$ variables requires the inversion of only two matrices of orders p and q. The analysis is conveniently carried out in covariance form, and can easily be extended since the calculation of the over-all covariance matrix effectively brings us back to a single covariance and the process can be repeated again. In multiple regression terminology these formulae are equivalent to the statement that a regression of y on $x_1, \cdots, x_p, z_1, \cdots, z_q$, can be carried out by calculating initial regressions of $y, z_1, \cdots, z_q$ on $x_1, \cdots, x_p$ and subsequently a regression of the residuals of y on the residuals of $z_1, \cdots, z_q$ when $x_1, \cdots, x_p$ have been eliminated.

This analysis may be conveniently split up into the following calculations:

(1) the sums of squares and products of all variables,
(2) the inverse matrix $C(x_i x_j)$,
(3) the coefficients b_k and the matrix B,
(4) the matrices $S_x(z_i y)$ and $S_x(z_i z_j)$,
(5) the inverse matrix $C_x(z_i z_j)$,
(6) the coefficients d and e and the residual $\sum_{zz} y^2$, and
(7) the over-all covariance matrix.

Step (1) would be carried out whatever the method of analysis, while steps (3), (4), and (6) are roughly equivalent to the multiplication that would be required after the calculation of an over-all inverse matrix. Thus the above calculation effectively replaces the inversion of an over-all matrix by steps (2), (5), and (7), and the inversion of, say, a 12 by 12 matrix can be replaced by the inversion of two 6 by 6 matrices together with three matrix multiplications.

TABLE IV

TEST OF THE DIFFERENCE BETWEEN MALE AND FEMALE ADOLESCENT CONSUMPTION

	Degrees of Freedom	Sum of Products of ξ and s	Sum of Products of l and s	Sum of Products of q and s
Mean	1	−0.500	−11.250	56.333
Workers	3	0.833	− 1.417	4.500
$w + u - 2v$	1	0.050	0.200	0.092
Residual	55	−5.383	26.467	−12.925
	60	−5.000	14.000	48.000

		Sum of Products of C and s	Sum of Squares of s
Mean	1	74953.2	0.41667
Workers	3	−1252.5	2.75000
$w + u - 2v$	1	158.0	0.00833
Residual	55	−2232.7	21.82500
	60	71626.0	25.00000

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 0.0291991 & 0.0008544 & 0.0009425 \\ 0.0008544 & 0.0016646 & 0.0000218 \\ 0.0009425 & 0.0000218 & 0.0013459 \end{bmatrix} \begin{bmatrix} -5.383 \\ 26.467 \\ -12.925 \end{bmatrix} = \begin{bmatrix} -0.146758 \\ 0.039175 \\ -0.021892 \end{bmatrix}$$

Residual sum of products of C and s (corrected for ξ, l, and q) equals $-2232.7 - 1009.3 = -3242.0$.
Residual sum of squares of s (corrected for ξ, l, and q) equals $21.82500 - 2.10983 = 19.71517$.
Regression coefficient d is $-3242.0/19.71517 = -0.0507224(3242.0) = -164.4$.

$$-0.0507224\ \mathrm{B} = \begin{bmatrix} 0.0074439 \\ -0.0019870 \\ 0.0011104 \end{bmatrix}, 164.4\ \mathrm{B} = \begin{bmatrix} -24.1 \\ 6.4 \\ -3.6 \end{bmatrix}, e = \begin{bmatrix} 1304.4 \\ 251.1 \\ -133.9 \end{bmatrix}$$

$$0.0507224\ \mathrm{B'\ B} = \begin{bmatrix} 0.0010925 & -0.0002916 & 0.0001630 \\ -0.0002916 & 0.0000778 & -0.0000435 \\ 0.0001630 & -0.0000435 & 0.0000243 \end{bmatrix}$$

Thus the covariance matrix for e_1, e_2, e_3, and d is

$$\begin{bmatrix} 0.0302916 & 0.0005628 & 0.0011055 & 0.0074439 \\ 0.0005628 & 0.0017424 & -0.0000217 & -0.0019870 \\ 0.0011055 & -0.0000217 & 0.0013702 & 0.0011104 \\ 0.0074439 & -0.0019870 & 0.0011104 & 0.0507224 \end{bmatrix}.$$

In practice the inversion of anything greater than an 8 by 8 matrix is tedious. The 3 by 3 matrix would seem to be a convenient unit with which to work, although when we are uncertain whether certain factors should be taken into account these may be tested and, if necessary, incorporated.

A further point that should be noted is that the set of orthogonal factors eliminated by the analysis of variance can be considered as the first step in the above process and that, correspondingly, these factors can be incorporated into the over-all inverse matrix at any stage.

Example. To demonstrate the form of this analysis, we might test whether the accuracy of the analysis carried out in Tables II and III might be improved by differentiating between the sexes of children over 13

and whether a reliable estimate of any such differences can be obtained. This is done using a dummy variate ξ which takes values $+1$ for male adolescents, -1 for female adolescents, and 0 for other children. The form of the analysis is shown in Table IV. The estimated regression coefficient in this case is -164.4 ± 421.9 so there is no gain from introducing this extra variable, the accuracy of which is not very good. However, it has been assumed that this extra variable has been taken into account and an over-all covariance matrix has been calculated.

One further point should be noted about the form of analysis: the individual analyses of variance reveal wherever two sets of factors in the collected data deviate very greatly from orthogonal samples and therefore indicate where the existence of interaction will be of greatest importance.

4. DETERMINATION OF THE WEIGHTS

It has already been pointed out that an increase in size of family normally leads to an increase in the variation and that for the method of least squares to be efficient the observations should be weighted according to their relative accuracy. This can be done according to some predetermined scale or according to a rough initial analysis. Fortunately it has been shown [**4**] that the analysis is accurate even if the weighting is fairly rough. However some care is necessary in carrying out the over-all analysis because of the problems raised by interaction with size of family. Thus, for example, it is fairly obvious that the effect of location and economic classification will vary with size and constitution of family. Normally the assumption that these effects are constant for families of a particular size will be sufficiently accurate, but any further assumption concerning their variation with size of family will lead to a greater degree of inaccuracy in the representation. Again, if the individual consumption varies with size of family this tends to give an incorrect picture in an over-all analysis. Thus, for example, if the relative consumption decreases with size of family, since the larger, i.e., older, families contain a greater proportion of adolescents, this effect, if not taken into account, will be shown in an over-all analysis by a decrease in the estimated consumption of the adolescents.[4] Thus the following approach is suggested.

[4] There are, however, two other points which should be noted at this stage. Firstly, if interaction is detected in analyses by size of family, the combination of these analyses cannot be carried out uncritically if the results are to be of general application. Thus, if children in small families are getting more than their requirements and children in large families less than their requirements, to state that the average is "just right" hardly does justice to the situation, while a standard error attached to such an average acquires a specialized meaning. Secondly, an effect of the kind described above can occur irrespective of size of family if there is a tendency towards "arrested development" in the family consumption.

As a first step in the analysis we eliminate effects that are known to vary with size of family. A rough analysis is then carried out eliminating the effect of family constitution. This is most conveniently carried out using graduated parameters, as in the above example. Unless great accuracy is required, the type of work need not be taken into account since the determination of the weights to within 20% will usually suffice, although if it is desired to investigate the extent and form of interaction larger analyses will be required. The relative weights given by the reciprocals of the residual mean squares can now be used to combine the individual analyses *after* the stage where the effects varying with size of family have been eliminated. The analysis then follows the usual lines with the initial matrix $C(x_i x_j)$ now taking the form:

$$\begin{bmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & C_n \end{bmatrix}.$$

However, it may subsequently be necessary to combine or analyse the estimates obtained by size of family, so that some general remarks on the combination of least-squares estimates are necessary.

5. THE COMBINATION OF LEAST-SQUARES ESTIMATES

If we have a series of normally and independently distributed estimates $\theta_1, \theta_2, \cdots, \theta_n$, of a parameter θ, with variances $V_1, V_2, \cdots, V_n$, then it is well-known that the most efficient combined estimate $\bar{\theta}$ is given by $(\sum V_i^{-1})\bar{\theta} = \sum V_i^{-1}\theta_i$, with a variance $\bar{V}$ given by $\bar{V}^{-1} = \sum V_i^{-1}$. Similarly, if we have a series of normally and independently distributed estimates $\theta_1, \theta_2, \cdots, \theta_n$, of a set of parameters θ, with covariance matrices $V_1, V_2, \cdots, V_n$, then the most efficient combined estimates are given by $(\sum V_i^{-1})\bar{\theta} = \sum V_i^{-1}\theta_i$, with a covariance matrix $\bar{V}$ given by $\bar{V}^{-1} = \sum V_i^{-1}$. This, in effect, states that an over-all least-squares analysis should be carried out with weighting proportional to the residual mean squares in the individual analyses. An alternative approach to this seems worth noting. If we have two sets of estimates θ_1, θ_2 with covariance matrices V_1, V_2, then we can find a transformation A such that $A'V_1A$ and $A'V_2A$ are diagonal and, correspondingly, so that θ is transformed to $A'\theta$. The transformed set of estimates are independent of each other so that the ordinary weighting rules will apply and

$$[(A'V_1A)^{-1} + (A'V_2A)^{-1}]A'\bar{\theta} = (A'V_1A)^{-1}A'\theta_1 + (A'V_2A)^{-1}A'\theta_2,$$

i.e.,

$$A^{-1}(V_1^{-1} + V_2^{-1})\bar{\theta} = A^{-1}(V_1^{-1}\theta_1 + V_2^{-1}\theta_2),$$

which is the same formula as obtained previously. This alternative ap-

proach, however, shows that the loss of information resulting from the direct combination of least-squares estimates will depend upon the deviations of V_1 and V_2 from diagonal matrices. In particular, the deviation of $V_1^{-1}V_2$ or $V_2^{-1}V_1$ will largely determine the loss of accuracy for the combination of matrices. Also it suggests that an initial transformation to make the matrices V_1 and V_2 approximately diagonal might be fairly efficient.

As a first step in investigating the efficiency of direct combination we might consider least-squares matrices of the kind $A_1P_1A_1$ where A_1 is diagonal and

$$P_1 = \begin{bmatrix} 1 & p_1 & p_1 \cdots p_1 \\ p_1 & 1 & p_1 \cdots p_1 \\ p_1 & p_1 & 1 \cdots p_1 \\ \cdots & \cdots & \cdots \\ p_1 & p_1 & p_1 \cdots 1 \end{bmatrix}.$$

This type of matrix will arise when each observable has an equal probability of occurrence with every other observable. The inverse of this matrix is $A_1^{-1}P_1^{-1}A_1^{-1}$, where

$$P_1^{-1} = \begin{bmatrix} P_1 & Q_1 & Q_1 \cdots Q_1 \\ Q_1 & P_1 & Q_1 \cdots Q_1 \\ Q_1 & Q_1 & P_1 \cdots Q_1 \\ \cdots & \cdots & \cdots \\ Q_1 & Q_1 & Q_1 \cdots P_1 \end{bmatrix},$$

$$P_1 = \frac{1 + p_1(n-2)}{(1-p_1)[1 + p_1(n-1)]},$$

$$Q_1 = -\frac{p_1}{(1-p_1)[1 + p_1(n-1)]},$$

and n is the order of the matrix. This matrix can now be used to investigate the loss of information. Thus, if $A_1 = a_1^{\frac{1}{2}}I$, so that the least-squares matrix is a_1P_1, and the estimates from this matrix are directly combined with those from a second matrix a_2P_2, then the percentage loss of information is found to be

$$\frac{100(n-1)(p_1-p_2)^2}{(1-p)\{1+p(n-1)\}\{1+p_1(n-2)\}\{1+p_2(n-2)\}} \cdot \frac{a_1a_2}{(a_1+a_2)^2},$$

where

$$p = \frac{a_1p_1 + a_2p_2}{a_1 + a_2}.$$

The latter half of this expression depends only upon the relative

number of observations in the two groups and cannot take a value exceeding $\frac{1}{4}$ which it does when $a_1 = a_2$. Table V gives the values of this expression for $a_1 = a_2, 2a_2$; $n = 3, 5, 7, 9$; $p_1 = 0.0, 0.2, 1.0$; $p_2 = 0.0, 0.2, 1.0$. It is apparent that the loss of information is serious only when the difference between p_1 and p_2 is large, i.e., in effect, when one set of observations might be used to offset a high correlation between estimates obtained from the other. This effect is naturally more important if the

TABLE V

PERCENTAGE LOSS OF INFORMATION RESULTING FROM THE DIRECT COMBINATION OF LEAST-SQUARES ESTIMATES

	$a_1 = a_2$						$a_1 = 2a_2$					
p_1 / p_2	0.0	0.2	0.4	0.6	0.8	1.0	0.0	0.2	0.4	0.6	0.8	1.0
n = 3												
0.0	0.0	1.5	5.1	10.0	16.5	25.0	0.0	1.3	4.5	9.3	16.4	28.6
0.2	1.5	0.0	1.1	3.9	8.3	15.1	1.4	0.0	1.0	3.6	11.4	18.0
0.4	5.1	1.1	0.0	0.9	3.6	8.9	4.6	0.9	0.0	0.8	4.7	11.0
0.6	10.0	3.9	0.9	0.0	0.9	4.8	8.9	3.4	0.8	0.0	0.9	6.2
0.8	16.5	8.3	3.6	0.9	0.0	1.9	14.1	6.9	2.9	0.8	0.0	2.6
1.0	25.0	15.1	8.9	4.8	1.9	0.0	20.0	11.5	6.5	3.4	1.4	0.0
n = 5												
0.0	0.0	2.0	5.0	8.3	12.1	16.7	0.0	1.7	4.3	7.3	11.4	18.2
0.2	2.0	0.0	0.7	2.3	4.4	7.3	1.9	0.0	0.6	2.1	4.3	8.5
0.4	5.0	0.7	0.0	0.4	1.6	3.6	4.9	0.7	0.0	0.4	1.6	4.3
0.6	8.3	2.3	0.4	0.0	0.4	1.7	7.9	2.0	0.4	0.0	0.4	2.1
0.8	12.1	4.4	1.6	0.4	0.0	0.6	11.0	3.8	1.3	0.3	0.0	0.8
1.0	16.7	7.3	3.6	1.7	0.6	0.0	14.4	5.8	2.7	1.2	0.4	0.0
n = 7												
0.0	0.0	2.1	4.5	6.9	9.4	12.5	0.0	1.7	3.7	5.9	8.7	13.3
0.2	2.1	0.0	0.5	1.5	2.7	4.3	2.0	0.0	0.4	1.3	2.6	4.9
0.4	4.5	0.5	0.0	0.2	0.9	1.9	4.5	0.5	0.0	0.2	0.9	2.3
0.6	6.9	1.5	0.2	0.0	0.2	0.9	6.8	1.3	0.2	0.0	0.2	1.1
0.8	9.4	2.7	0.9	0.2	0.0	0.2	8.9	2.5	0.7	0.2	0.0	0.4
1.0	12.5	4.3	1.9	0.9	0.2	0.0	11.1	3.5	1.4	0.6	0.2	0.0
n = 9												
0.0	0.0	2.1	4.0	5.8	7.7	10.0	0.0	1.7	3.2	4.9	7.0	10.5
0.2	2.1	0.0	0.4	1.0	1.8	2.9	2.1	0.0	0.3	0.9	1.7	3.2
0.4	4.0	0.4	0.0	0.2	0.5	1.2	4.2	0.3	0.0	0.1	0.5	1.4
0.6	5.8	1.0	0.2	0.0	0.1	0.5	5.9	0.9	0.1	0.0	0.1	0.6
0.8	7.7	1.8	0.5	0.1	0.0	0.2	7.5	1.6	0.5	0.1	0.0	0.2
1.0	10.0	2.9	1.2	0.5	0.2	0.0	9.1	2.3	0.9	0.4	0.1	0.0

highly correlated estimates involve a greater proportion of the total information available, and it would appear to be generally true of all least-squares estimates that they can be combined without much loss of information provided there is not an appreciable change in the interaction between the factors involved in each estimate. For example, the direct combination of estimates obtained from the least-squares matrices

$$\begin{bmatrix} 1 & 1-\epsilon & 0 & 0 \\ 1-\epsilon & 1 & 0 & 0 \\ 0 & 0 & 1 & 1-\epsilon \\ 0 & 0 & 1-\epsilon & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 1-\epsilon & 0 \\ 0 & 1 & 0 & 1-\epsilon \\ 1-\epsilon & 0 & 1 & 0 \\ 0 & 1-\epsilon & 0 & 1 \end{bmatrix},$$

and

$$\begin{bmatrix} 1 & 0 & 0 & 1-\epsilon \\ 0 & 1 & 1-\epsilon & 0 \\ 0 & 1-\epsilon & 1 & 0 \\ 1-\epsilon & 0 & 0 & 1 \end{bmatrix},$$

where ϵ is small, would involve almost complete loss of information since the high correlations may be used to offset each other.

Thus, while the direct combination of least-squares estimates is feasible, it requires a careful consideration of whether the individual groups of observations may augment one another, and here again the possibility of using a transformation presents itself.

6. SURVEY PLANNING

It has been suggested above that, from the point of view of the statistical analyses, orthogonality or partial orthogonality is a property to be desired in survey planning. These suggestions are worthy of further discussion in relation to the execution of the survey and the application of its results. Naturally, if we wish to investigate particular points, emphasis should be laid upon these points to ensure that accurate comparisons can be made. For example, if it is desired to estimate the effect of size of family on individual consumption, a higher proportion of large families than would normally occur might be included in the survey to ensure that this comparison is as accurate as possible. Again, the joint and independent estimation of two or more effects requires the use of orthogonal samples, and the existence of interactions is best recognised with orthogonal samples. However, apart from the difficulties of collecting such a sample, other difficulties will often make it impracticable. Over-all comparisons will normally be made by weighting estimates from different groups according to the proportions in some standard population which is used as a basis of comparison, and if this results in a relatively large loss of information then some other method must be used. Thus the process of taking an orthogonal sample might lead to an imbalance in the proportions of families of different sizes or in the proportions of children of different ages. Also, an orthogonal sample will not necessarily measure what is required whereas this might be provided by a nonorthogonal sample. For example, consider an investigation in which we have two economic classifications, "rich" and "poor," and two family size classifications, "large" and "small"; the virtual nonexistence of the "large-rich" groups would necessitate a great deal of extra sampling if the two comparisons are to be made orthogonal. However, in fact the comparison of the dietary practices of large and small

families must inevitably be referred to the "poor" group if the comparison is to be a valid one, and correspondingly in determining the effect of income group small families must be used. Hence the extra sampling works to the disadvantage of the final result.

This rather baldly leaves out of account finer points such as the interaction of the two factors both from the statistical point of view and in consequence of the maxim that a rise in family size causes a relative decrease in income, but it is fairly obvious that orthogonality is not always to be desired. Nevertheless, when the sets of factors involved scarcely interact, if an orthogonal sample can be collected without undue difficulty the subsequent analysis might be greatly simplified. Similarly, if the population gives rise to almost orthogonal samples, then orthogonal samples might be taken and subsequently adjusted for the effect of interactions on the over-all comparisons that are to be made. Larger departures from orthogonality might require, if interaction is present, a partially-orthogonal sampling design.[5] There are a large number of such designs depending upon various partitions of sums of squares which can be carried out for any or all of the methods of classification, so that it will usually be possible to find a design which would be reasonably representative. However, if the method of weighting suggested above is adopted, then the factors whose effects vary with size of family, i.e., economic, geographical, and seasonal effects, must be eliminated first of all. Thus the sample should be designed to allow these factors to be eliminated simultaneously with a minimum of trouble. Only when this can be done will attention to the other factors be justified.

So far nothing has been said about the method of sampling families of different sizes and the best methods of estimating the individual consumption of children of different ages. Theoretically these can be studied by fractional replication designs, but from the practical viewpoint this is both tedious and wasteful in relation to the comparisons that will finally be made. In practice, therefore, we shall with certain reservations usually be content with stratifying for other factors, including possibly size of family, and analysing the results by the analysis of covariance. The reservations arise in consequence of the interactions of family constitution with other factors, which may cause the almost complete identification or confounding of two or more effects if special action is not taken. Thus, for example, special action may be necessary to prevent the effect of size of family being identified with an increase in age and manifested by an apparent decrease in adolescent consumption.

Hence, to summarise, the sample should deviate from a random or stratified random sample only in so far as it is necessary to ensure that

[5] If a sample is taken so that two or more effects are nearly orthogonal, then it is often possible to treat the sample as if it were orthogonal without undue qualms.

particular effects are distinguished and, if possible, to allow economic, geographical, and seasonal effects to be eliminated simultaneously.

University of Aberdeen

REFERENCES

[1] S. W. CLEMENTS, "A Family Coefficient Scale Developed from the Australian Nutrition Survey," *Journal of Hygiene*, Vol. 40, December, 1940, pp. 681–689.

[2] M. H. QUENOUILLE, "The Analysis of Covariance and Nonorthogonal Comparisons," *Biometrics*, Vol. 4, December, 1948, pp. 240–246.

[3] W. G. COCHRAN, "The Omission or Addition of an Independent Variate in Multiple Linear Regression," *Supplement to the Journal of the Royal Statistical Society*, Vol. 5, No. 2, 1938, pp. 171–176.

[4] J. W. TUKEY, "Approximate Weights," *Annals of Mathematical Statistics*, Vol. 19, March, 1948, pp. 91–92.