Should Weighted or Unweighted Data be Used in Analysis?

It isn't always easy to decide. In simple analysis, data must be weighted nearly always. Suppose one has a frequency distribution of personal incomes: to calculate average income, clearly one must use the formula $\Sigma\ y_i n_i/N$ and *not* $\Sigma\ y_i/K$, where y_i is the average income in the ith income group, K number of classes.

If, however, one is making, say,a household budget study with many factors giving e.g. number of persons in household, social status of head of household, household income etc. and one wishes to explain average expenditure on a particular item in terms of these factors it may be misleading to use data for *all* families in the country (if one has the data!) in one's multivariate regression calculations. Or, at least, one must understand what happens in so doing. If the country has many poor and few rich, what will emerge (coefficient-wise etc is relationship applying virtually to the poor, which may or may not apply to the rich. Clearly the ideal would be to study each income group separately and compare the results in each group. In each group unweighted data are used, i.e. each family is given the weight unity. Of course, the same procedure would apply to a large random sample.

At the other extreme, if one has data for only two families, say a rich family and a poor family, each deemed typical,it is perfectly meaningful, on simple analysis,to point out that the Engels ratio for food for the rich family is 20 *per cent* compared with 50 *per cent* for the poor family. The fact that there are many poor and few rich is irrelevant.

At a more sophisticated level, suppose that one has average income and the Engels ratio for food for each income group and no other data of family characteristics,it is meaningful to regress the ratio on average income. Having found the regression coefficient b (with a negative sign) what one is saying is that a rise of £1 in income entails a lowering of b *per cent* in the ratio. That is to say,one contemplates a single family moving up £1 in the income scale. Here, also, the *number* of families on the scale before and after is irrelevant.

The use of unweighted data amongst one's independents in multivariate regression leads to a very simple result which may be well-known but which was not known to the writer. This result is that the independents are mutually orthogonal!

Let there be two independents X_1 and X_2 and let there be, say, three values of X_1 namely X_{11}, X_{12}, X_{13} and, say, four values of X_2, namely X_{21}, X_{22}, X_{23}, X_{24}. Let Y_{ij} $(i = 1, 2, 3; \quad j = 1, 2, 3, 4)$ be the value of the dependent variable corresponding to X_{1i}, X_{2j}. There are N = 12 sets of observations. The sum product m is then

$$m = \Sigma_i \Sigma_j X_{1i} X_{2j} - 12\, \overline{X}_1 \overline{X}_2 = \Sigma_i X_{1i} \Sigma X_{2j} - 12\, \overline{X}_1 \overline{X}_2$$

with

$$12\overline{X_1} = 4\Sigma X_{1j}; \qquad 12\overline{X}_2 = 3\Sigma X_{2j}$$

On substitution, m = 0. Obviously the result is general, as regards number of factors and numbers of classes.

This fact is without objective significance. Suppose two of the independents are income and social status (measurable in numerical terms). Objectively there is a high positive correlation between the measures which can easily be calculated. (Actually from the Irish 1965-66 Inquiry r = , using 1, 2, 3, 4, 5 the descending order numbers as measures of social grade (non-agricultural).

The result (m = 0) is of great convenience, since it means that the multivariate regression can be broken down into a series of simple regressions. The calculated value Y_c is the algebraic sum of the contribution of each factor, no matter how many factors there are.

8 December, 1969. R.C. Geary