Relationships Between Group Averages and Individual Observations

By Martin E. Abel and Frederick V. Waugh[1]

CORRELATION AND REGRESSION equations are often based upon group averages rather than upon individual observations. It is not always recognized that the results may be substantially different, especially the two kinds of correlation. This does not mean that one kind of correlation is always "right," nor that the other kind is "spurious" or "biased." The correlation that is appropriate in a particular case depends upon the purpose of the study.[2]

In many practical cases, the correlations based upon group averages tend to be higher than those based upon individual observations. An example of this was brought to the attention of one of the authors some years ago by the late Dr. Margaret Jarman Hagood, in connection with her work on levels of living. She computed the correlations in table 1 which show the relationships between total farm family living expenditures and several items. The first column in the table shows the correlations based upon 3,985 individual observations. We shall use the cooefficient $r_{(ind)}$ throughout this paper to indicate the correlations based upon individual observations. The second column is based upon the same data as column 1, but the data were averaged for each of 97 counties. Throughout the paper we will use the coefficient $r_{(gr)}$ to represent correlations based upon group averages.

Note that in all cases the correlations based upon group averages are higher than those based upon individual observations. The correlations

Table 1.--Correlations between total farm family living expenditures and several items, 1955

Item	$r_{(ind)}$ based upon 3,985 individual observations	$r_{(gr)}$ based upon 97 averages by counties
Value of products	0.33	0.73
Education expense	.43	.50
Medical care expense	.56	.79
Recreation expense	.14	.82
Reading expense	-.11	.82

in the second column are what the sociologists commonly call ecological correlations; meaning correlations based upon averages by counties, States, or other geographical areas. Ordinarily these ecological correlations and many other correlations based upon group averages tend to be substantially higher than the correlations based upon individual observations. This fact was pointed out by Robinson (7). Goodman (2, p. 611) states that "it has been shown that ecological correlations cannot be used as substitutes for individual correlations. However, ecological correlations may be of interest in themselves; the kinds of questions that can be answered by a study of ecological correlations are sometimes of direct concern to social scientists. In some problems, both ecological and individual correlations and the relations between them may be of interest. Even if the investigator is concerned only with the individual correlations, ecological data may be of service, though ecological correlations are not recommended."

Recently Grunfeld and Griliches (6) discussed the more general problem of relationships based upon group averages. They presented

[1] The authors wish to express their appreciation to Clark Edwards, Economic Research Service, whose assistance greatly improved the presentation in this paper, particularly in the last section.

[2] A similar point about correlations between ratios is made by F. E. A. Briggs (1, p. 162). (Underscored numbers in parentheses refer to the Literature Cited, p. 115.)

two interesting, practical examples, together with a mathematical analysis which indicates that we should expect that a correlation based upon group averages would be higher than a correlation based upon the individual observations from which the group averages were derived.

We propose in this paper to analyze a fairly simple statistical problem. We have used this particular example to compute a number of regression equations and correlation coefficients. In each case we have computed two regressions and two correlations: one based upon individual observations, and another based upon group averages.[3] We shall first present the results of these computations. Then we shall give a brief theoretical analysis of the problem.[4] And finally, we shall discuss some of the implications of these findings for economic research.

The Example to be Analyzed

We have illustrated this problem by an analysis of data concerning the income and value of food consumed by 80 low-income families in a large Eastern city. These families were all receiving some kind of public assistance. Some of them were receiving cash welfare payments, some were receiving donated surplus foods, and some were receiving both forms of assistance. The data were gathered by the U.S. Department of Agriculture in the early spring of 1961 for use in connection with the development of pilot food stamp programs in eight areas of the country.

The data obtained from these 80 families included information on the value of food consumed during a one-week period, on family income (including relief payments) in the previous month, on the number of persons in the family, and on several other matters that do not concern us here.

RELATIONSHIPS BASED UPON INDIVIDUAL OBSERVATIONS

Figure 1 is a typical "dot chart." Each of the 80 solid dots shows the monthly income, X, and food consumption (in value terms), Y, by one of the individual families. The solid line marked (1) is the regression of food consumption upon family income; i.e., it is the line to be used in estimating the expected food consumption of an individual family associated with a given level of family income. It shows that an increase of one dollar in family income is associated with an average increase of $0.105 in value of food consumed. The dashed line marked (2) is the regression of family income upon food consumption; i.e., it is the line to be used in estimating the expected income of an individual family associated with a given level of food consumption. It shows that an increase of one dollar in food consumption is associated with an increase of $2.252 in income. Both of these regressions were computed on the basis of the 80 individual observations. The squared correlation r^2_{xy} is equal to the product of the two regression coefficients; i.e.,

(1) $$r^2_{xy} = 2.252 \times 0.105 = 0.24.$$

You could see that the correlation is small without computing the correlation coefficient. The smallness of the correlation is indicated by the wide scatter of the dots in figure 1.

RELATIONSHIPS BASED UPON AVERAGES BY INTERVALS OF INCOME

The solid dots in figure 2 are the same as those in figure 1. But in figure 2 we have classified the families into seven income groups: those getting less than $100 a month, those getting from $100 to $125, etc. The circled dots represent the average family income and average food expenditures in each income group. The two regression lines in figure 2 are both regressions of Y on X; that is, both lines are estimates of expected food consumption associated with given amounts of family income. The solid line

---

[3] An example of this type of comparison is found in George R. Rockwell, Jr. (8).

[4] Readers are referred to the Grunfeld-Griliches article (6) for a more rigorous mathematical treatment of the problem.

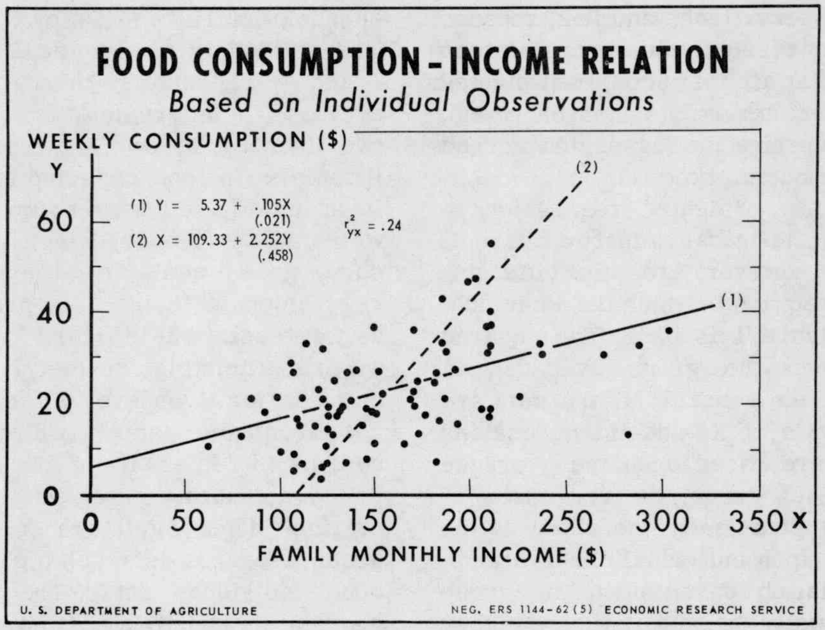

Figure 1

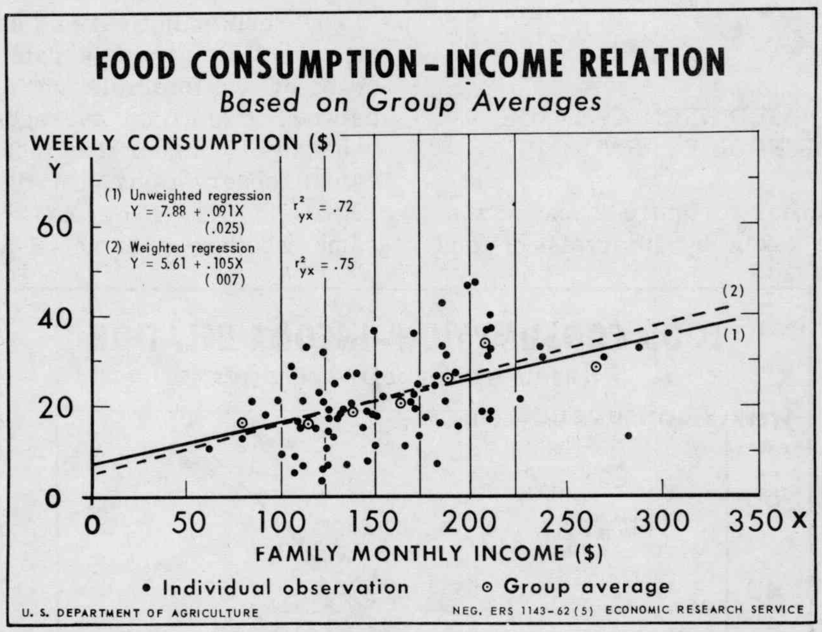

Figure 2

marked (1) is an unweighted regression; that is, it gives each group average the same weight regardless of the fact that group 1 includes only 5 families, whereas group 2 includes 21 families, for example. The dashed line marked (2) is a weighted regression; that is, it weights each group average by the number in the group.

In many cases when statisticians have used group averages as the basis for regressions and correlations, they have not bothered to weight the averages by the number of observations. This, we believe, is a reprehensible practice because it gives averages based on a few observations the same weight as averages based on many observations, and it introduces a bias in the regression equation. If group averages are used at all, they should be weighted by the number of observations in each group. Thus, we strongly prefer the regression marked (2) rather than the one marked (1).

We note that the weighted regression in figure 2 is almost identical to that in figure 1. The correlations, however, are quite different. The squared correlation between the individual observations in figure 1 is 0.24. The squared correlation between the group averages in figure 2 is 0.75. In general, if the data are grouped by intervals of X, and if regressions and correlations are based upon the averages of X and Y in each group, the regression of Y upon X will be practically the same as the regression based upon individual observations; while the correlation based upon the group averages will usually be substantially higher than that based upon individual observations. We will discuss the reasons for this later in the paper.

RELATIONSHIPS BASED UPON AVERAGES BY INTERVALS OF FOOD EXPENDITURES

Figure 3 is similar to figure 2 except that the averages are taken by intervals of food consumption rather than by intervals of income. The solid dots are identical to those in figure 1 and 2. The circled dots represent the group averages by intervals of food consumption. The two lines in figure 3 are regressions of family income upon food consumption. Again, line (1) is an unweighted regression, and line (2) is a weighted regression, which we prefer. In this case, lines 1 and 2 are almost identical. Either regression in figure 2 is practically the same as regression (2) in figure 1. In other words, if we are estimating expected family income associated with a given level of food consumption, we get about the same results from the group averages by intervals of food expenditures as we do by computing a regression based upon the original data. Again, the correlation between group averages is much higher than that based upon individual data. The weighted squared correlation based upon group averages in figure 3 is 0.78 compared with the squared correlation of 0.24 based upon the individual observations.

The results in figures 2 and 3 are typical. In general, if we group the data either by intervals of X or by intervals of Y, the correlations between the group averages will usually be substantially higher than those between the individual observations. On the other hand, the regression of Y on X is usually about the same whether it is based upon the individual

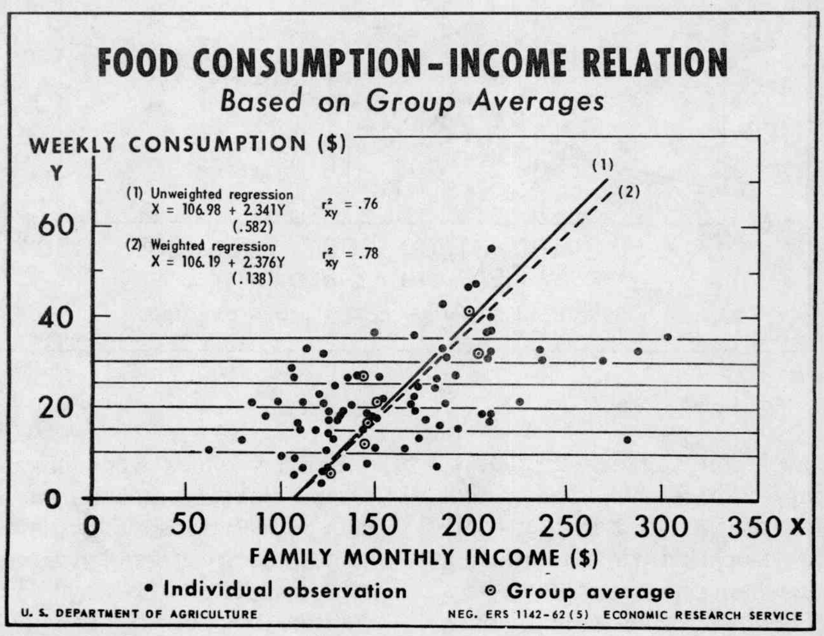

Figure 3

observations or upon averages of intervals of X. In a similar way, the regression of X on Y is about the same, whether it is based upon individual observations or upon group averages by intervals of Y.

Before leaving figures 2 and 3, we would like to make two general comments. First, we think it is always desirable to draw dot charts and to indicate group averages on the charts. In this way the researcher can see the nature of the data and can decide, for example, whether to fit a linear function or some type of curve. Methods of doing this have been discussed by Ezekiel (3, pp. 431-453) and also by Ezekiel and Fox (4, ch. 14). Second, we would emphasize that in any problem of two variables, say X and Y, there are always two regressions: the regression of Y upon X, and the regression of X upon Y. Each of these regressions has a definite meaning. If the researcher wants to estimate the expected values of Y associated with given values of X, he should group the data by intervals of X and vice versa.

RELATIONSHIPS BASED UPON TWO-WAY CLASSIFICATION

In figure 4 the data are classified into four intervals of family income, and each of these intervals is subclassified into four intervals of food consumption. In all, there are 16 subgroups or "cells." The circled dots indicate the averages of the items in each cell. We computed two regressions, using the cell averages as the observations. The solid line labeled (1) is the weighted regression of food expenditures upon family income. The dashed line labeled (2) is the weighted regression of family income upon food expenditures. Both of these lines in figure 4 are approximately the same as the corresponding lines in figure 1. For that reason the correlation is about the same. The squared correlation based upon the group averages in figure 4 is 0.30 compared with 0.24 in the case of the correlation based upon individual observations in figure 1.

When dealing with large numbers of observations, such as Census schedules for individual farms in the United States for example, we think it is often desirable to make a two-way classification like figure 4, and to work out group averages. This is somewhat similar to the "correlation tables" which were used so effectively by Yule and Kendall (9, ch. 11) to illustrate a wide variety of correlation and regression problems. The main difference is that Yule and Kendall simply counted the different observations in each cell and assumed that the group averages were at the centers of the cells.

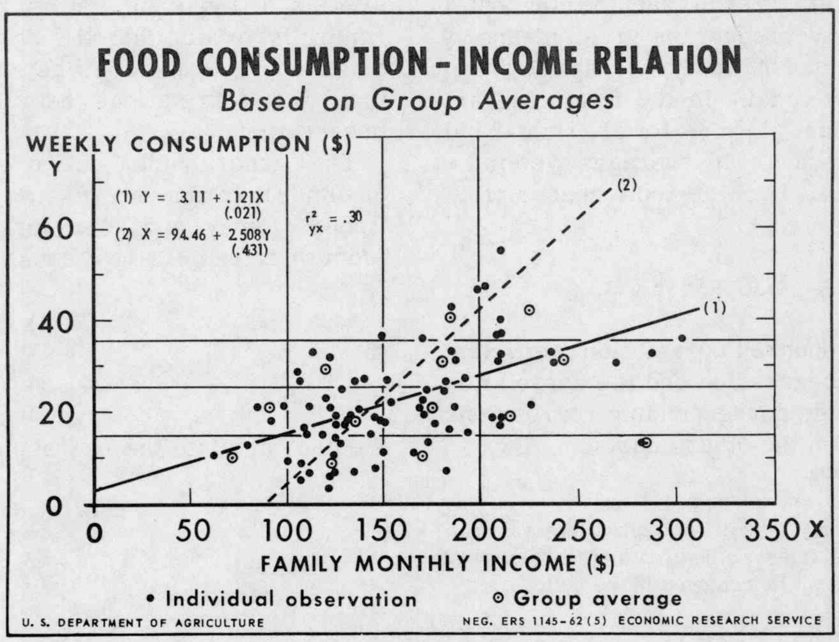

Figure 4

In the cases discussed by Yule and Kendall this made no significant difference. But, in principle, we think it is preferable to compute the group averages[5] as we have done and to show them in a dot chart like figure 4. When dealing with as few observations as in our example, this can make a substantial difference.

RELATIONSHIPS BASED UPON CLASS INTERVALS OF ANOTHER VARIABLE

Instead of classifying the data by intervals of family income or by intervals of food consumption, we might classify them by some other variable, such as size of family. This kind of problem often confronts any statistician who works with data that are grouped according to intervals of such factors as occupation, size of business, or geographical location. It is thus one of the problems underlying the so-called ecological correlations of the sociologist.

This problem is a little difficult to illustrate on a small chart. Instead we have taken the same data as those shown in figures 1 and 4, and have shown them in a series of diagrams in figure 5. For example, the little diagram in the top center shows only the dots for families of two persons. The diagram in the upper right-hand corner is a similar dot chart for families of three persons, etc. In each case, the horizontal and vertical lines represent the group means of food consumption and of family size. These group means are shown again in the final diagram which can be found in the lower right-hand corner. Using these group averages we find a weighted regression of food consumption on family income

$$Y = -14.30 + 0.15X \tag{2}$$

and the weighted squared correlation coefficient is 0.97. Both the regression and the correlation are higher than the corresponding regressions and correlations on the other charts.

[5] This is particularly true if wide class intervals are used. The wider the interval the greater the possibility that the average will not be centered in the cell.

Before giving a technical explanation for differences in regressions and correlations, we note that the main reason is that in this case family size is very highly correlated both with family income and with food consumption. These families were all getting some kind of public assistance, and the payments were based in large part upon family size. Those who were getting surplus foods also received amounts of food that were related to family size. This suggests that whenever we are interested in the relation between two variables, and use group averages based upon intervals of another variable, the results may be very different from the relationships based upon individual observations, especially in any case when the data are classified by another variable that is closely associated with either X or Y. For example, if we use State averages to study the relation between family income and food consumption, our results would be affected by the fact that there are substantial differences in average family income in different States.

An Explanation

We have already indicated that the results found in the particular example are fairly typical. When the data are grouped in ways similar to those in figures 2, 3, and 5, we can generally expect that the correlations between the group averages will be substantially higher than the correlations between the individual observations.

The general idea is very simple. Let us assume that the regression of Y upon X is linear. Let y and x be deviations of Y and X from their respective means. Then

$$\begin{aligned} y &= (Y-\overline{Y}) \\ x &= (X-\overline{X}). \end{aligned} \tag{3}$$

We can calculate the regression lines

$$\begin{aligned} y &= b_{yx}x \\ x &= b_{xy}y; \end{aligned} \tag{4}$$

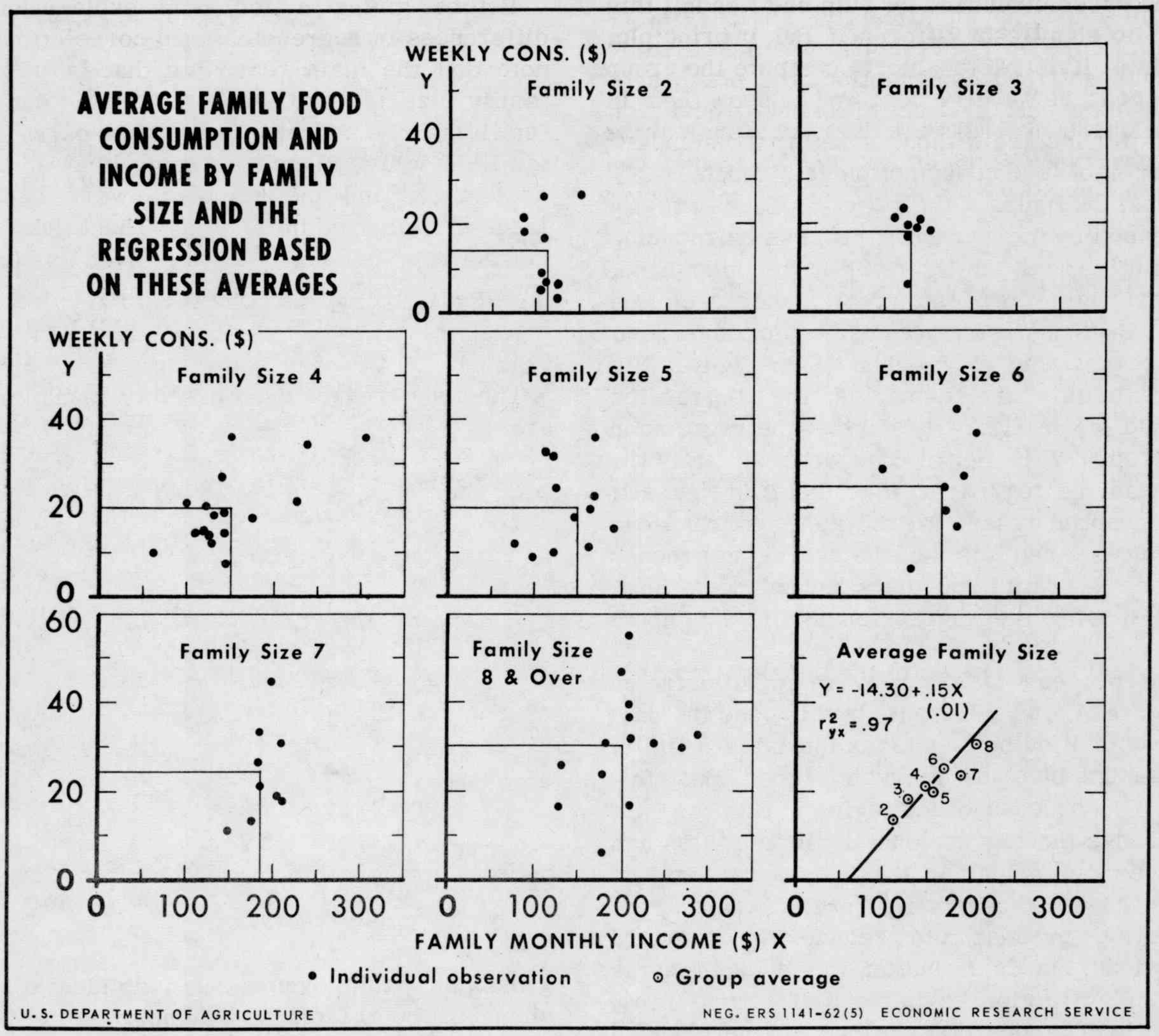

Figure 5

the standard deviations of y and x, S_y and S_x; and the correlations between the two variables

$$r_{yx} = b_{yx} \frac{S_x}{S_y}$$

(5)

$$r_{xy} = b_{xy} \frac{S_y}{S_x}.$$

To understand the difference between correlations based upon group averages and those based upon individual observations, we must center our attention on the standard deviations, S_x and S_y. The process of averaging by groups always reduces these standard deviations--but usually not in the same proportions. This can be seen clearly in figures 2 and 3.

For example, compare the standard deviations of the group averages shown in figure 2 with the standard deviations of the individual observations. The process of averaging by intervals of family income reduced the standard deviations of food consumption very substantially, while it did not change the standard deviation of family income significantly. The reverse is true in figure 3. When the data are classified by intervals of food consumption, the process of averaging greatly reduces the standard deviation of family income without changing the standard deviation of food consumption very much.

Thus, in figure 2 the ratio, S_x/S_y, is greatly increased by the averaging process by reducing S_y relative to S_x. In figure 3, on the other hand, the averaging process greatly increases the ratio, S_y/S_x. We have already pointed out that the regression of X upon Y in figure 3 is approximately the same as the regression marked (1) in figure 1, and that the regression of X upon Y in figure 3 is approximately the same as the regression marked (2) in figure 1. The essential reason for the higher correlations in figures 2 and 3 is that the averaging process raised the ratios of the standard deviations, while leaving the regression coefficients about the same.

The process of group averaging illustrated in figure 4 was different. In this case the data were classified both by class intervals of family income and by class intervals of food consumption. The process of averaging in this case did not change the regressions significantly, as can be seen by comparing the two regressions in figure 4 with those in figure 1. Of course, the averaging process did reduce the standard deviations, but the reduction was about the same in the X direction as in the Y direction. Thus, the ratio of standard deviations was not changed significantly. With approximately the same regressions and the same ratio of standard deviations, we get approximately the same correlation coefficients--in this case with group averages as with individual observations.

We can demonstrate mathematically the effects of group averages on the regression and correlation coefficient. Let x_1 and y_1 denote the means of the various groups of x and y, and x_2 and y_2 be the deviations of the individual x's and y's from their respective group means. Thus,

$$x = x_1 + x_2 \qquad (6)$$
$$y = y_1 + y_2 .$$

We proceed to compute the individual regressions, $b_{yx(ind)}$ and $b_{xy(ind)}$. These are

$$b_{yx(ind)} = \frac{\Sigma yx}{\Sigma x^2} = \frac{\Sigma x_1 y_1 + \Sigma x_2 y_2}{\Sigma x_1^2 + \Sigma x_2^2} \qquad (7)$$
$$b_{xy(ind)} = \frac{\Sigma xy}{\Sigma y^2} = \frac{\Sigma x_1 y_1 + \Sigma x_2 y_2}{\Sigma y_1^2 + \Sigma y_2^2}$$

where

$$\Sigma x_1 x_2 = \Sigma y_1 y_2 = \Sigma x_1 y_2 = \Sigma x_2 y_1 = 0, \qquad (8)$$

exactly.

The group regressions based on group means are

$$b_{yx(gr)} = \frac{\Sigma y_1 x_1}{x_1^2} \qquad (9)$$
$$b_{xy(gr)} = \frac{\Sigma x_1 y_1}{\Sigma y_1^2}.$$

We can call $\frac{\Sigma y_1 x_1}{\Sigma x_1^2}$ and $\frac{\Sigma x_1 y_1}{\Sigma y_1^2}$ the regressions

between groups and $\frac{\Sigma y_2 x_2}{\Sigma x_2^2}$ and $\frac{\Sigma x_2 y_2}{\Sigma y_2^2}$ the regressions within groups. The individual regressions will be greater than, equal to, or less than the group regressions depending upon whether the regressions within groups are greater than, equal to, or less than the values of the regressions between groups. That is,

$$b_{yx(ind)} \gtreqless b_{yx(gr)} \text{ if } \frac{\Sigma y_2 x_2}{\Sigma x_2^2} \gtreqless \frac{\Sigma y_1 x_1}{\Sigma x_1^2} \qquad (10)$$
$$b_{xy(ind)} \gtreqless b_{xy(gr)} \text{ if } \frac{\Sigma x_2 y_2}{\Sigma y_2^2} \gtreqless \frac{\Sigma x_1 y_1}{\Sigma y_1^2}.$$

In many cases, we would expect the slopes of the regressions within groups to be about the same as those between groups. In such cases, the regressions based on individual observations would be about equal to the regressions based on group averages. This was true in our example when we grouped by intervals of X, by intervals of Y, and by intervals of X and Y.

When we grouped by size of family, however, the regressions of Y on X within groups were substantially smaller than that between groups. For that reason the individual regression was significantly less than the group regression.

The individual squared correlation is

$$r^2_{xy(ind)} = b^2_{yx(ind)} \frac{\Sigma x^2}{\Sigma y^2} = b^2_{yx(ind)} \frac{\Sigma x_1^2 + \Sigma x_2^2}{\Sigma y_1^2 + \Sigma y_2^2}$$

$$(11) \qquad = b^2_{xy(ind)} \frac{\Sigma y^2}{\Sigma x^2} = b^2_{xy(ind)} \frac{\Sigma y_1^2 + \Sigma y_2^2}{\Sigma x_1^2 + \Sigma x_2^2}$$

while the group correlation is

$$(12) \quad r^2_{xy(gr)} = b^2_{yx(gr)} \frac{\Sigma x_1^2}{\Sigma y_1^2} = b^2_{xy(gr)} \frac{\Sigma y_1^2}{\Sigma x_1^2}.$$

Thus, the major difference between $r_{xy(ind)}$ and $r_{xy(gr)}$ is in the ratios of variances of X to those of Y. As we have shown in our example, these ratios are often changed markedly by grouping, and in a predictable direction.

From (11) and (12) it follows that the relationship between the group and individual correlation coefficients will be

$$r^2_{xy(gr)} \gtreqless r^2_{xy(ind)}, \text{ or}$$

$$(13) \quad b^2_{yx(gr)} \frac{\Sigma x_1^2}{\Sigma y_1^2} \gtreqless b^2_{yx(ind)} \frac{\Sigma x^2}{\Sigma y^2}, \text{ or}$$

$$\frac{b^2_{yx(gr)}}{b^2_{yx(ind)}} \gtreqless \frac{\dfrac{\Sigma x^2}{\Sigma y^2}}{\dfrac{\Sigma x_1^2}{\Sigma y_1^2}}.$$

In general we would expect the correlation coefficient based on group averages to be larger than the one based on individual observations. However, it does not always have to be this way. A simple example shown in figure 6 will suffice to illustrate the reverse situation. The example is constructed so that the slope of the regression line within each group is the same as the one between groups. Each individual observation and each group average are the same vertical distance from the regression line. The process of averaging reduces the variance of X more than it reduces the variance of Y. The regression coefficient remains unchanged. It follows, then, from (13) that $r^2_{yx(gr)}$ would be less than $r^2_{yx(ind)}$. One, of course, could also illustrate this same point with an example that appeared less extreme.

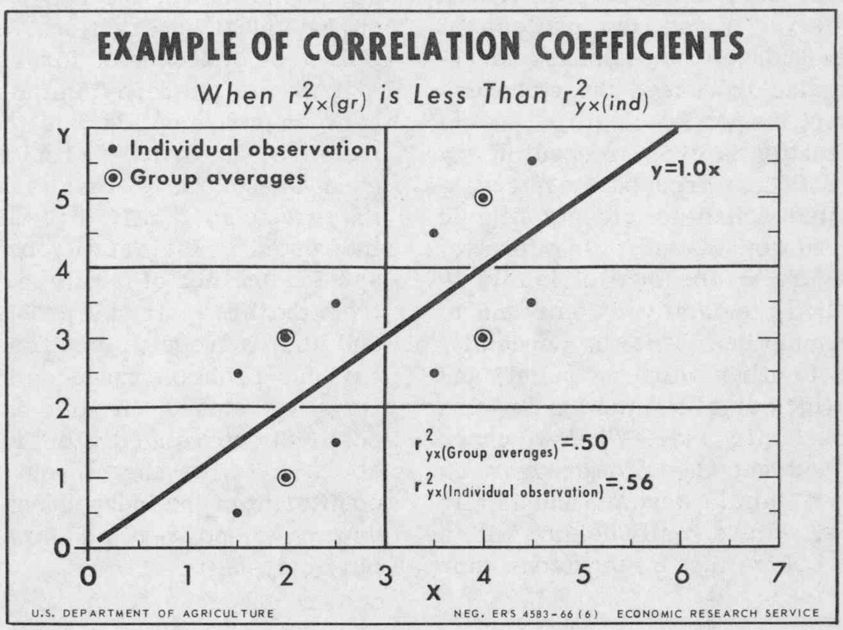

Figure 6

One purpose of using group averages is to reduce the labor of computation. This apparently was one of the main purposes of Yule and Kendall. Even in these days of automatic computation, there may be significant savings in time and cost by using group averages. In addition, graphical presentations of the data are simplified when group averages are used. And, if our purpose is to estimate the regressions and correlations based upon the individual observations, we recommend a two-way grouping, such as that used so effectively by Yule and Kendall, and such as we have illustrated in figure 4.

Of course, some judgment is needed in determining the class intervals in each direction. We suggest that there should be roughly the same number of class intervals in either direction, and that the intervals be so chosen as to give a reasonable number of observations in each cell.

The situation illustrated by figure 5 is a little different from those in figures 2, 3, and 4. In the case illustrated by figure 5, the data were first grouped by size of family. For each family size we computed averages of family income and family food expenditures. The relation between these group averages is shown in the bottom right-hand corner.

In this case the squared correlation was raised to 0.97 (compared to 0.24 in the case of individual observations). The correlation was raised in two ways. First, the process of averaging greatly reduced the standard deviations. Second, it also increased the steepness of the regression of Y upon X.

This kind of situation is likely to occur in any case where the data are grouped by intervals of a third variable which is closely related to X, to Y, or to both X and Y. In our case, illustrated by figure 5, the size of family is closely related both to family income and to family food consumption. This is especially true since these families were on relief, and relief payments (including donated foods) were based partly upon family size. When we group by family size and get the relation between group averages of family income and family food expenditures, we are really dealing with a three-variable problem; that is, food consumption is affected not only by family income, but also by family size. The regression line shown in the lower right diagram in figure 5 and the associated correlation coefficient actually combine the effects of family income and family size upon food consumption.

The data presented in figure 5 were used in a three-variable problem. The estimated regression equation is

$$Y = 4.28 + \underset{(0.026)}{0.066X} + \underset{(0.547)}{1.379Z} \tag{14}$$

where Z is family size. The squared multiple correlation coefficient is 0.29. The squared simple correlation coefficient between Y and Z is 0.23 and between X and Y is 0.36. However, the squared simple correlation between group averages of X and Y, classified by size of family, is 0.97. It can readily be seen that widely different results are obtained when the data are grouped by intervals of a third variable and when the third variable is used in a multiple regression and correlation problem. Notice also that the coefficient of income is lower than in the regression of consumption on income based on individual observations. This is not unreasonable since in the three-variable problem the coefficient of X is net of changes in family size, whereas in the two-variable case based on individual observations, the effect of family size was not accounted for.

If we are interested in the gross relationship between consumption and income, the results from the two-variable problem are appropriate, even though these results are due in part to the effects of family size. If, however, we are interested in the relation between consumption and income net of family size, then the results from the three-variable problem are appropriate and the regression coefficient from the two-variable problem based on individual observations is biased. If the omitted variable is positively correlated with the independent variable, as is the case in our problem, then the coefficient of the independent variable is biased upward (5, pp. 8-20). This we readily see from our analysis.

Literature Cited

(1) Briggs, F. E. A. The influence of errors on the correlation of ratios. Econometrica, Vol. 30, No. 1, January 1962.

(2) Goodman, Leo A. Some alternatives to ecological correlation. Amer. Jour. Sociol., Vol. LXIV, No. 6, May 1959.

(3) Ezekiel, Mordecai. A method of handling curvilinear correlation for any number of variables. Jour. Amer. Statis. Assoc., Vol. XIX, Dec. 1924.

(4) Ezekiel, Mordecai, and Karl A. Fox. Methods of correlation and regression analysis. Wiley & Sons, New York, 3d ed., 1959.

(5) Griliches, Zvi. Specification bias in estimation of production. Jour. Farm Econ., Vol. XXXIX, No. 1, February 1957.

(6) Grunfeld, Yehuda, and Zvi Griliches. Is aggregation necessarily bad? Rev. Econ. and Statis., Vol. XLII, No. 1, February 1960.

(7) Robinson, W. S. Ecological correlations and the behavior of individuals. Amer. Sociol. Rev., Vol. 15, No. 3, June 1950.

(8) Rockwell, George R., Jr. Income and household size: Their effects on food consumption. U.S. Dept. Agr., Mktg. Res. Rpt. 340, June 1959.

(9) Yule, G. Udney, and M. G. Kendall. An introduction to the theory of statistics. Griffen & Co., London, 13th ed., 1945.