Letter to the Editor

Dear Mr. Mattingly:

Someone must write a little more precisely but somewhat less gently than did Mr. Katz about the techniques that Messrs. Denton and George used in their "Socio-Economic Influences on School Attendance: A Study of a Canadian County in 1871." Katz, in his earlier article, had used two- and three-way tables in order to separate out the effects on school—attendance of a variety of influences—child's age, family size, economic status, ethnic status as measured by birthplace of parent, and so forth. They fault this procedure for not attaching precise measures of association or influence to the several variables. And, of course, the procedure does not do this. As an improvement and refinement, they then use multiple regression on a rather similar population. They describe this procedure as one by which the separate influences of variables can be "taken into account and controlled for." Here they err. They imply, for example, that their regression coefficients and the associated t-ratios provide a way of answering the question: Controlling school-attendance for the obvious effect on it of children's age (since older children attend less), what is the influence on attendance exerted by other considerations such as ethnicity or family size? But multiple regression doesn't do that. Let me put this as untechnically as I can. Multiple regression takes all the variance in something being studied (school attendance), and establishes coefficients for an equation to estimate that something from a collection of independent variables. *If* these independent variables are also independent of each other, then the regression coefficients do provide an estimate of the relative contribution that these independent variables make to the whole estimate. *But*:

1. If much of the variance is accounted for by some variable that is not relevant to the question being asked (here, age of child), the measures of significance still assess the other variables for how much they contribute to the over-all variance, including that accounted for by age. They do *not,* as long as the investigators introduce no control-for-level, provide any answer to, say, how much socio-economic factors account for school-attendance patterns at any particular level of age. Denton and George do not introduce such a control, and the technique does not automatically embody such a control. To put it crudely: the variance within any one age-level may be fairly small, to be studied by fine measuring instruments; but they have inflated the whole process by demanding that non-age variables help to account for the gross fact that older children drop out of school. I suppose that this is a subtle form of statistical fudging, but such amounts of sugar and chocolate are still unhealthy.

2. If the variables entered into a multiple regression are not independent of each other, then the relative sizes and even signs of the regression coefficients cannot be interpreted as estimates of the relative importance of variables. Several such misleading effects do appear in the Denton-George data. For example, high-status occupations and urban residence are (at least for many populations) fairly highly correlated. Yet the coefficients for high-status occupations are positive, while the coefficient for urban residence is negative. These differences cannot be interpreted as if they were analogous to partial correlation coefficients. Quite often, if two independent variables are highly correlated, then the machinery of multiple regression will

attribute to them large but opposite-signed coefficients. It is not an accident, in the Denton-George data, that all the signs for father's-birthplace coefficients are opposite to the corresponding signs for mother's-birthplace coefficients. Fortunately the authors do not make much attempt to build argument on the particulars of their coefficients, but the little that they do make is too much.

A more satisfactory procedure for their study might have been:

a. to use age as one separate variable in their analysis
b. to perform a principal-components factor analysis on the other independent variables, across the whole population, omitting age from this analysis to construct the table of inter-correlations for all resulting variables:
   school-attendance, age, and the several principal components (these perhaps rotated), and then
d. to deflate the whole age variable out of the correlation table.

The results of all this might well still show that occupational or urban/rural variables are more important than demography or ethnic variables, but the point would then be plainly established.

By the way, I must beg off from some responsibility that Mr. Katz would lay on me. His reference to my Dutchess County data in his Reply implies that I had presented material with the same rigor essayed by Messrs. Denton and George (or by him). I said that the Dutchess County data "suggest" certain relationships between urban family size and school attendance. To claim more, I would have to take the time to follow my own advice.

Daniel H. Calhoun
University of California,
Davis