DISCUSSION

Halliman H. Winsborough, University of Wisconsin, Madison

In a number of places, Fienberg and Mason observe that the identification problem in cohort analysis pertains to the linear component of the effects of age, period and cohort on the dependent variable. Although this fact is relatively obvious once one has seen it, arriving at an understanding of the point has caused me some pain. It is perhaps useful, therefore, to discuss an alternative approach to the observation.

The identification problem was, in recent times at least, posed for demographers and sociologists by Ryder who observed that if the cohort variable is expressed as year of birth, age in years, and period by a date, then period equals age plus cohort. Thus, it does not make any sense to run a regression on, say, fertility rates of the form:

$$F = a+b_1A+b_2P+b_3C+e$$

where all variables are thought to be continuous, the reason being, of course, that the X'X matrix is not of full rank. The whole business of using dummy variables and more recently log-linear analysis for such problems has been an attempt to deal with this difficulty and with the fact that one would not expect the effects of age, period, or cohort to be linear anyway.

Given the latter observation, it is useful to investigate what would happen if we permitted nonlinear effects in the continuous variables model. Suppose, for example, we thought that

$$Y = a+b_1A+b_2P+b_3C+B_4A^2+b_5P^2+b_6C^2+e.$$

Here it is still the case that P = A+C. Thus, we will not be able to separate the three linear terms. We could, however, estimate

$$Y = a+(b_1+b_2)A+(b_3+b_2)C+b_4A^2+b_5P^2+b_6C^2+e$$

by substituting the identity for P. We have no difficulty estimating b_4, b_5 and b_6 since the linear identity yields

$$P^2 = A^2+2AC+C^2$$

and substitution gives

$$Y=a+(b_1+b_2)A+(b_3+b_2)C+(b_4+b_5)A^2+(b_6+b_5)C^2+2b_5AC+e.$$

Thus, substitution of the squared identity yields an equation in only A and C from which b_4, b_5 and b_6 are retrievable. A similar retrieval is possible for higher-order polynomials in P, A and C.

Insofar, therefore, as one generates a cohort model of the form

$$Y = f(P)+g(A)+h(C)+e$$

and insofar as f, g and h can usefully be approximated by a Taylor series, the identification problem exists only with the linear terms in each series. Setting any one of the three coefficients to zero will allow estimation of all higher-order terms.

What does this fact imply for exploratory cohort analysis, i.e., analysis of a substantive problem for which it is not possible to make a strong a priori assertion about the process--assertions such as Fienberg and Mason are able to creditably make with their education example? It implies that the second and all higher derivatives of f, g and h can be uniquely estimated but the first derivative cannot. (This is the rationale behind the somewhat cryptic observation in the Fienberg - Mason paper that the "acceleration" of a variable with age, period, or cohort is estimable.) An analogue of this assertion holds for discrete coding of the independent variables. For the dummy variable method of analysis, Mason et al. [1] present an example showing that various choices of identifying restriction yield quite different effect parameters. Inspection of that example shows that the first differences of these parameters also vary with identifying restrictions. But second and all higher-order differences are invariant.

If one knows rather little about the process he is investigating, this estimability of second and higher-order differences in effect parameters is, I suppose, better than nothing. It certainly implies that such cohort models are not "hopelessly" under-identified, if that sometimes emotional phrase is meant to suggest that exploratory cohort analysis is inherently doomed to be completely unproductive.

There is a second point about the estimability problem in cohort analysis which seems to me important to make. The point is that the problem is not inevitable. It is not inevitable because Ryder's identity that period equals age plus cohort is not true for many data structures. Consider respondents born in the year 1900 and interviewed on July 1, 1976. If these respondents are asked their age, about half of them will correctly respond that they are 75, while the rest will correctly declare they are age 76. These people born on or before July 1, 1900 are 76; those born after July 1, 1900 are 75. This fact is, of course, familiar to demographers and is represented in that field by the Lexis diagram. Thus, if one's model posits a set of effects associated with number of whole-life-years lived and a set associated with period of interview, and if interviews do not occur on the first of January, the Ryder identity simply does not hold. Period will, of course, have quite a high multiple correlation with age and cohort but the resultant multi-colinearity is conceptually rather different than the estimability problem.

It can be argued that the above assertion represents a kind of "trick." If variables were scaled continuously, and thus accurately represented the "real" continuous nature of time, it can be argued, the identity would hold. There are two reasons why such an assertion of the greater reality of continuity is as "tricky" as my assertion. First, at the limit, it is difficult to believe that birth is an instantaneous

process. It seems more reasonable to regard birth as a process occupying a time interval with any assignment of days, hours, minutes and seconds to the event as a crude way of locating the interval in time. Second, many of the dependent variables of interest represent some measure of a process which occurs in an interval. Examples are whether or not a child was born in the last year, earnings in the last year, or hours worked last week. Thus, it seems to me the argument that the Lexis diagram escape from the trap is a trick is based on an argument about the greater "reality" of continuity which is itself a trick. Of course, taking advantage of the Lexis diagram to ameliorate the problem requires that the analyst have great control over his data. Particularly it requires information which can yield both year of birth and current age. For many problems in the analysis of archival data such control is not available. In these situations, the estimability problem will exist.

REFERENCES

[1] Mason, Karen Oppenheim and William M. Mason, H. H. Winsborough, W. Kenneth Poole. "Some Methodological Issues in Cohort Analysis of Archival Data." American Sociological Review 38 (April): 242-258. 1973.