COHORT ANALYSIS IN SOCIAL RESEARCH

Beyond the Identification Problem

Edited by
William M. Mason · Stephen E. Fienberg

With 44 Figures

Springer-Verlag
New York Berlin Heidelberg Tokyo

and two additional papers, one by Robert Johnson and one by James Heckman and Richard Robb that address issues not already covered in the conference presentations. This set of papers thus represents a substantial updating of material presented in the original conference. In particular, the identification problem in its simplest form is a point of departure for most of the chapters. Indeed, most of these contributions offer researchers assistance with a variety of problems in cohort analysis that do not relate directly to the identification problem.

The appearance of these papers as a published volume is the result of a major collaborative effort involving substantial contributions from Burton Singer, Peter Read and Herbert Smith, as well as those of the editors. We are pleased to acknowledge here the persistence, commitment and hard work of these three colleagues. The volume itself was produced using computing and typesetting facilities at Carnegie-Mellon University. Diana Bajzek was an invaluable aid in preparing the copy for typesetting, and for the generous lending of her expertise and guidance to this project we are most grateful. Barbara Krest prepared the computerized text for the entire manuscript, while Margie Krest prepared the mathematics and the tables, supervised the proofreading, and was responsible for the final aspects of photo-typesetting. Without all of their outstanding efforts and assistance we would have been unable to complete the volume.

William M. Mason
Stephen E. Fienberg

CONTENTS

5. USING LONGITUDINAL DATA TO ESTIMATE AGE, PERIOD AND COHORT EFFECTS IN EARNINGS EQUATIONS*

James Heckman

Richard Robb†

The literature on the determinants of earnings suggests an earnings function for individual i which depends on age a_i, year t, "vintage" or "cohort" c_i, schooling level s_i, and experience e_i. Adopting a linear function to facilitate exposition we may write

$$Y_i(t,a_i,c_i,e_i,s_i) = \alpha_0 + \alpha_1 a_i + \alpha_2 t + \alpha_3 e_i + \alpha_4 s_i + \alpha_5 c_i , \qquad (1)$$

where e_i is experience, usually defined for males as age minus schooling, $(e_i = a_i - s_i)$,[1] and Y_i may be any monotone transformation of earnings.

Each variable in equation (1) has some argument supporting its inclusion in the equation. Age (a_i) may be a direct determinant of earnings through maturation or other physiological effects. It may also be a signal used by employers to estimate productivity. Year effects (t) may arise from disembodied technical progress that affects all workers or else through other general labor market variables that determine individual earnings. Work experience (e_i) is interpreted as a proxy for "human capital acquisition" (see Mincer, 1974) or other activities that raise the perceived productivity of workers as a function of their work experience.

* This research was supported by NSF Grant SES-8107963, NIH Grant NIH-1-R01-H016846-01 and a Guggenheim Fellowship awarded in 1978–1979. The first draft of this paper was read at an SSRC Conference at Mr. Kisco, New York, October, 1978. We have benefited from discussions with Tom MaCurdy. A later draft was presented at an LSE Conference on The Analysis of Panel Data on Income in June, 1982.

† The authors are affiliated with the Department of Economics, University of Chicago, and the National Opinion Research Center.

[1] This measure is less adequate for females. See Mincer and Polachek (1974) for measures of work experience for females that recognize the intermittent nature of lifetime female labor supply.

Schooling (s_i) is known to increase earnings (but the precise reason for this is not yet known). The cohort variable (c_i) is a stand in for variety of plausible cohort specific phenomena. Easterlin (1978) and Welch (1979) suggest that workers born into a large cohort carry an economic disadvantage throughout their careers (and conversely for those born into small cohorts). Easterlin (1961) suggests that workers of one cohort have different labor market expectations and therefore may pursue different careers and family plans (e.g., workers reared in the Depression may be pessimistic or risk averse while workers reared in the 1950s may be overly optimistic). A variety of other cohort specific phenomena might be suggested. These arguments taken together justify the list of variables used in equation (1) but not necessarily the functional form of the equation.[2]

It is obviously not possible to estimate all of the coefficients of equation (1) without further restrictions because of the definitional interdependencies among the variables. That is,

$$e_i = a_i - s_i ,$$

$$t = a_i + c_i .$$

Direct estimation of (1) by least squares is impossible because of multicollinearity among the variables.

One can substitute for c_i and a_i to reach a commonly utilized specification of the earnings function:

$$Y_i(t,e_i,s_i) = [\alpha_0 + (\alpha_2 + \alpha_5)t] + [\alpha_1 + \alpha_3 - \alpha_5]e_i + [\alpha_1 + \alpha_4 - \alpha_5]s_i . \qquad (2)$$

In a cross section, the year effect $[(\alpha_2 + \alpha_5)t]$ is impounded in the intercept term. Clearly one can estimate the *difference* between the effect of experience (α_3) and the effect of schooling (α_4) in a cross section. But holding schooling constant, an extra year of experience is also an extra year of age as well as reduction in the person's (birth) cohort by one year.

[2] We consider more general functional forms in the next section.

A similar difficulty arises in estimating the schooling coefficient holding experience constant.

The way we have defined them, age, period, and cohort variables and experience, age and schooling variables obey exact linear relationships among each other. This suggests that equation (1) makes no sense. An analyst cannot claim that the variables on the right hand side of (1) can be independently varied. Equation (1) is not identified.

The justification for (1) that implicitly appears in the literature is that most or all of the variables on the right hand side of (1) are proxy variables for underlying unobserved variables that are not themselves linearly dependent. Thus age (a_i) is a proxy for physiological variables and screening measures used by employers. The year effect (t) is a proxy for macroeconomic or market variables that in principle can be measured. The experience variable (e_i) is a proxy for training and other variables that raise perceived productivity that in principle can be measured. Schooling (s_i) is only a proxy for ability or productivity (or both) and the cohort variable (c_i) is a crude proxy for a variety of causal variables.

This argument suggests that equation (1) is only a crude approximation of what we are really interested in. The approximation is so crude that it creates a problem of its own. The correct formulation of (1) is as an errors in variables model (see, e.g., Aigner *et al.* (1984) for a survey of such models). Equations like (1) should not be taken too literally and at a minimum one should look for better proxies for the underlying unobserved variables.

This paper considers the general issue of whether or not longitudinal data can be used to estimate equation (1) or its errors in variables analogue. We conclude that the age-period-cohort problem as currently formulated is ill posed. On the issue of errors in variables, we do not have much to say beyond pointing to the existing econometric literature that indicates how access to longitudinal data affords the analyst additional instrumental variables in estimating an errors in variables model. But this point is worth making because this approach is not currently used in age-period-cohort analysis.

We make the following additional points. (1) Contrary to Fienberg and Mason (1978), the linear dependencies in the definitions of variables of equation (1) affect identification of interaction and higher order terms (Fienberg and Mason, 1985, correct this mistake in their paper in this volume). (2) Convenient "normalizations" in fact are often interpreted as if they have content. They may lead to erroneous interpretations of the data. (3) Additional information must be used to break the identification problem.

1. THE IDENTIFICATION PROBLEM

The central issue considered here is whether or not longitudinal data (repeated observations over time on the same individuals) can be used to identify further parameters of equation (1) above and beyond the estimable combinations of parameters in equation (2) that can be identified from a cross section. It is by now well known (e.g., Cagan, 1965) that such data do not solve the identification problem. The "year effect"—previously impounded in the intercept in equation (2) operates in the course of a panel, unless the economic environment is stationary.

Suppose that the year effect can be ignored ($\alpha_2 = 0$). In this case, longitudinal data can be used to estimate some of the parameters of the model. For example, two successive post school observations on the same cohort permit estimation of ($\alpha_1 + \alpha_3$), an age plus experience effect. Then, clearly the effect of cohort, α_5, can be estimated from the cross section (see equation (2) and the discussion surrounding it). Hence $\alpha_1 + \alpha_4$ can be estimated. Note that panel data and a time series of cross sections of unrelated individuals are equally informative on these parameters. Note further that as long as $\alpha_1 \neq 0$ all of the parameters of the model cannot be uniquely estimated.

The problem is that it is often plausible that $\alpha_2 \neq 0$. Suppose that we falsely assume that $\alpha_2 = 0$. Two successive (annual) observations on the same cohort provide an estimate of age plus experience plus year effects ($\alpha_1 + \alpha_2 + \alpha_3$). Assuming the year effect (α_2) is positive, $\alpha_1 + \alpha_3$ would be overestimated. In a cross section, earlier cohorts have both higher ages and experience levels. Thus for early cohorts, too large a component of earnings would be attributed to age plus experience. Too little would be attributed to the cohort effect. Hence the estimate of α_5, the cohort effect, would be downward biased.

Our discussion of identifiability also applies if each variable on the right hand side of (1) is measured as a categorical variable. Thus define

the vectors of dummy variables **A**, **T**, **E**, **S** and C to correspond to each of the continuous variables a, t, e, s, and c. **A** is a vector such that for a person age a, the ath element is 1 and all the remaining elements are zero. Each of the other vectors is defined correspondingly. Defining coefficient vectors D_j appropriately, and normalizing the first element in each vector to zero to avoid trivial linear dependencies associated with using dummy variables (see, e.g., Goldberger, 1968), we reach

$$Y_i(A_i,T,E_i,S_i,C_i) = D_0 + D_1A_i + D_2T + D_3E_i + D_4S_i + D_5C_i \,. \quad (3)$$

Corresponding to the condition $e_i = a_i - s_i$ in equation (1) is the condition for equation (3) that for an observation with the eth element of E_i nonzero, the jth element of A_i is nonzero if and only if the $(e-j)$th element of A_i is nonzero. (Since it makes no sense to have negative values of schooling, experience or age, it is necessary to require that $e \geq j$). Corresponding to the condition that $t_i = a_i + c_i$ in equation (1) is the requirement for equation (3) that for an observation with element t of T equal to one, the jth element of A_i is nonzero ($j \leq t$) if and only if the $(t-j)$th element of C_i is nonzero. Thus there are induced linear dependencies among the vectors of categorical variables and the D_i are not identified without further restrictions.

The remarks concerning identifiability made for the models with linear variables and with dummy variables carry over with full force to models with higher order interactions. Linear dependencies that arise from accounting identities become nonlinear identities. This point is not noted in the literature (see, e.g., Fienberg and Mason, 1978, p. 23). To illustrate this point consider the simple linear model of equation (1) and suppose $\alpha_2 = \alpha_5 = 0$. Thus, there is one restriction: $a_i = e_i + s_i$. Suppose that we are interested in exploring second order interaction terms as well. Thus to the model of equation (1), we add six terms for all the quadratic interactions and six parameters.

$$E_i(t,a_i,c_i,s_i) = \alpha_0 + \alpha_1 a_i + \alpha_3 e_i + \alpha_4 s_i + \beta_1 a_i^2 + \beta_2 e_i^2 + \beta_3 s_i^2$$
$$+ \beta_4 a_i e_i + \beta_5 a_i s_i + \beta_6 e_i s_i \,.$$

The identity $a_i = e_i + s_i$ implies the following three restrictions on the model:

$$a_i^2 = e_i^2 + 2e_is_i + s_i^2,$$

$$a_ie_i = e_i^2 + e_is_i,$$

$$a_is_i = e_is_i + s_i^2.$$

Clearly all three terms on the left-hand side of the equations must be substituted out to avoid exact linear dependencies in the model. Thus, from the six coefficients which arise when quadratic interaction terms are introduced, only three *combinations* of coefficients can be estimated. In a model with cubic terms, of the ten coefficients associated with the cubic terms, only four distinct combinations of coefficients can be estimated.

More generally, in a model with interaction terms of order k with j variables[3] and one linear restriction among the variables of the $\binom{j+k-1}{k}$ coefficients only $\binom{j+k-2}{k}$ *combinations* of the coefficients associated with terms of order k can be identified. Further, in a model with interaction terms of order k with j variables and l linearly independent restrictions on the $\binom{j+k-1}{k}$ coefficients, only $\binom{j+k-l-1}{k}$ *combinations* of the coefficients can be identified.[4] These propositions can easily be verified by elementary combinatorial algebra.[5]

2. THE DANGER OF TAKING A NORMALIZATION TOO LITERALLY

One approach to "solving" the age-period-cohort identification problem is to ignore one of the three effects. The danger in taking this approach

[3] An interaction term of order k is the product of k elementary variables. Thus $a_1^3a_2^4a_3^2$ is an interaction of order 9.

[4] Clearly $j + k - l - 1 > 0$ for the proposition to be meaningful.

[5] See, e.g., Feller (1968), Chapter 2, for one presentation of the relevant combinatorial calculations.

arises in confusing a normalization for a substantive interpretation. Studies by Weiss and Lillard (1978) and Johnson and Stafford (1974) provide an interesting illustration of this point. Both use NSF panel data on the earnings of scientists. Both find that for mathematicians and physicists the disparity in real wages between recent Ph.D. entrants and individuals with more professional experience increased during the 1960s. This is in contrast to the situation in economics and sociology where there is no evidence of increasing disparity between social scientists in the two experience classes.

Johnson and Stafford explain this result in terms of "year-experience interaction" in the labor market for scientists. Supplies of fresh Ph.D.'s were enlarged by federal subsidies just as demand for scientific personnel fell off. Due to a presumed rigidity in salary structures for tenured professors, much of the needed wage adjustment fell on young workers. A similar phenomenon was not at work in the market for social scientists where the demand for fresh Ph.D.'s held steady or even expanded.

The following modification of earnings equation (1) captures the Johnson-Stafford argument. Assume that there is no age effect and permit interaction terms between e and c, e and t and c and t but assume no interaction between s and the other variables. Then

$$Y(e,c,t,s) = \phi_0 + \phi_1e + \phi_2c + \phi_3t + \phi_4ec + \phi_5et + \phi_6ct + \phi_7s + \phi_8e^2 + \phi_9c^2 + \phi_{10}t^2 + \phi_{11}s^2. \tag{4}$$

The Johnson-Stafford story suggests that for scientists $\phi_5 > 0$, i.e., that in later years the contrast in earnings between experienced and inexperienced workers widens.

Weiss and Lillard claim that later cohorts have a higher growth rate of earnings due to "emobodied technical change." They argue that $\phi_4 > 0$. They offer evidence on this point from an empirical model that "normalizes" year effects to zero.

Using the definitions given above equation (2), $t = e + c + s$. Substitute into (4) to reach an equation similar to one utilized by Weiss and Lillard:

$$\begin{aligned}Y(e,c,s) &= \phi_0 + (\phi_1 + \phi_3)e + (\phi_2 + \phi_3)c \\ &+ (\phi_3 + \phi_7)s + (\phi_4 + \phi_5 + \phi_6 + 2\phi_{10})ec + (\phi_5 + 2\phi_{10})es \\ &+ (\phi_6 + 2\phi_{10})cs + (\phi_5 + \phi_8 + \phi_{10})e^2 + (\phi_6 + \phi_9 + \phi_{10})c^2 \\ &+ (\phi_{10} + \phi_{11})s^2.\end{aligned}$$

Evidence of a positive coefficient on the *ec* variable is clearly consistent with $\phi_4 = 0$, $\phi_5 > 0$ and $\phi_6 + \phi_5 + 2\phi_{10} > 0$. Thus the Weiss-Lillard evidence for their story is entirely consistent with the Johnson-Stafford story.

As there was no apparent "experience-cohort" interaction for social scientists, the Johnson-Stafford story appears to be the more plausible one for scientists because it is unlikely that embodiment of knowledge effects would be more pronounced for scientists than for social scientists.

This discussion illustrates two points: (1) The danger of taking a normalization too seriously, and (2) the value of additional information (e.g., the data on the market for social scientists) to resolve the identification problem.

3. SOLVING THE IDENTIFICATION PROBLEM WITH BETTER INFORMATION

The identification problem discussed in this paper arises because of dependencies among proxy variables. By assumption the underlying variables are not definitionally dependent. By improving the quality of the proxy variables it may be possible to secure identification of the effects of the unobserved variables and to avoid the collinearity that arises from using crude proxies.

The age-period-cohort effect identification problem arises because analysts want something for nothing: a general statistical decomposition of data without specific subject matter motivation underlying the decomposition. In a sense it is a blessing for social science that a purely statistical approach to the problem is bound to fail. We are forced to use our training as social scientists to improve on the crude age-period-cohort effect proxies (and the age, experience and schooling proxies).

One approach to this problem assumes that specific measured variables proxy the underlying unobserved variables. Thus in the context of an earnings equation, it is plausible to replace the year effect with variables indicating the state of the national and local labor market. Moreover, the concept of "cohort" can be refined. It is plausible that a cohort consists of a sequence of adjacent years (e.g., Depression or 1950s youth, etc.).[6]

If some individuals experience "breaks" in labor market experience so that $a_i \neq e_i + s_i$, and if the time spent in such breaks has no effect on earnings, it is possible to estimate separate age, experience and schooling effects in the earnings equation. The assumption that there is no parameter in the earnings function that is associated with time not at work is controversial. There are models such as those presented by Mincer and Polachek (1974) that give plausible reasons why time out of the labor force should result in skill atrophy.[7]

Other work by Lazear (1976) and Hanuschek and Quigley (1978) uses longitudinal data in an effort to break the exact linear dependence between age and experience. They measure experience by hours worked (see also Mincer (1978) who suggests an alternative to the Hanuschek-Quigley specification). It is important to note, however that human capital theory implies no simple or even montonic relationship between past hours of work and past investment, so that the value of this proxy is doubtful.[8]

[6] Cagan (1965) and Hall (1971) define vintage or cohort to be the same for an annual succession of vehicles based on comparable engineering specifications. Fienberg and Mason (1978) use a similar approach in their model of educational attainment.

[7] Even if there is no coefficient associated with time out of the labor force, age, experience and schooling are likely to be highly correlated so that estimates of these effects, if achieved, would not be expected to be very precise.

[8] Moreover, this "solution" creates the additional problem that work experience measures are not exogenous in an earnings equation.

4. A LATENT VARIABLE APPROACH

The approaches outlined in section **3** go only part way toward solving the problem. They achieve estimates of the "true" model only by assuming that some proxy is perfect. In fact, a better way to formulate the original problem is to write the earnings equation as a latent variable model.

To focus on essential ideas consider only age, period and cohort effects in a linear model. Earnings are assumed to be a function of a physiological variable at age a, P_a, a macro-variable at time t, M_t, and a cohort variable, E_c. Thus

$$Y(a,c,t) = P_a + E_c + M_t. \tag{5}$$

By assumption P_a, E_c and M_t are unobserved linearly independent variables. They may be decomposed into functions of observed and unobserved variables.

$$P_a = \phi_a(x_a\,;\theta_a) + u_a, \tag{6a}$$

$$E_c = \phi_c(x_c\,;\theta_c) + u_c, \tag{6b}$$

$$M_t = \phi_t(x_t\,;\theta_t) + u_t. \tag{6c}$$

The x_j, $j = a,c,t$ are observed exogenous variables. $E(u_j|x_j) = 0$. The x_j functional forms of ϕ_j are assumed known but the θ_j are unknown parameters. The covariance matrix for the u_i is unrestricted. Thus define $u = (u_a,u_c,u_t)$ and $\theta = (\theta_a,\theta_c,\theta_t)$, and

$$E(u'u) = \Sigma\,.$$

Inserting equations (6) into (5) results in a reduced form estimating equation for y in terms of the x variables. Standard identification criteria apply. If the ϕ functions are all linear, in order to be able to estimate θ no linear dependencies among the x variables are permitted. Access to panel data and/or repeated observations over time enables us to estimate the components of variance arising from u_a, u_c, and u_t.[9,10]

Proxies may also be available for P_a, E_c and M_t. The associated vectors of proxies are denoted respectively P_a^*, E_c^* and M_t^*. We assume that

$$P_a^* = b_a P_a + V_a, \tag{7a}$$

$$E_c^* = b_c E_c + V_c, \tag{7b}$$

$$M_t^* = b_t M_t + V_t, \tag{7c}$$

where V_i are suitably dimensioned mean zero vectors and the b_i are suitably dimensioned vectors. Intercepts could be added to these equations but this generality is foregone here to simplify the exposition. If the covariance matrices associated with V are diagonal we have the classical factor structure model provided that V is uncorrelated with P_a, E_c and M_t.

The conventional approach to age-period-cohort analysis assumes that a, c and t are perfect proxies for P_a, E_c, and M_t, respectively, except that the appropriate scaling elements (the b's) are unknown. Equivalently, these approaches assume that a, c, t are the sole determinants of P_a, E_c and M_t and that all of the ϕ functions are linear. The approaches surveyed in section 3 assume either that a richer (linearly independent) set of perfect proxies is available or that a richer set of x variables is available. The conventional approach makes overly strong assumptions and is wasteful of potential information. If we are prepared to be more specific about what our models mean and what we are trying to estimate, other proxies are available. Moreover, once we begin to take a position on what it is we seek to estimate we are usually able to produce x variables that plausibly determine P_a, E_c and M_t, so that the θ parameters can be estimated. With suitable covariance restrictions, equations (5)–(7) define a multiple-cause-multiple-indicator (MIMIC) model of the sort first analyzed by Jöreskog and Goldberger (1975) and extended in the econometric literature (see the survey by Aigner *et al.*, 1984).

There are a variety of plausible error structures for the u_j. For simplicity, we ignore covariance among different components of u. The most plausible error structure makes u_t identical across all observations at a common point in time but assumes u_t is serially correlated. There are two plausible specifications for u_c. The first assumes that u_c is common to all members of a cohort but lets u_c be serially correlated across cohorts. The second specification assumes that u_c is an independent draw from a

[9] The estimation problem is very similar to one considered by Nerlove (1967).

[10] Johnson and Hebein (1974) break the age-period-cohort effect identification problem by assuming that one of the components is an unobservable random component.

common distribution for each member of the cohort. The second specification assumes that u_c is an individual specific effect. Neither specification of u_c precludes the other.

Given the physiological interpretation placed on P_a and the known population variability in the aging process, it is plausible that u_a is a person specific component that varies across age for the same person. It is plausible that once the main effect of age is removed, the u_a components are uncorrelated across people but they may be correlated over age for the same person.

Panel data are required to identify the variance components of the person specific components u_a and u_c (under the second specification). A time series of cross sections suffices to identify the components of variances of u_a and u_c (under the first specification). A general analysis of identification in this model requires taking a position on the order of the time series process for each element of u as well as taking greater care in specifying the cross covariance terms in greater detail than has been done here. Such analysis is, by now, entirely conventional in econometrics. (See, e.g., Aigner *et al.*, 1984).

From the vantage point of the general MIMIC model presented in this section it should be clear that the linear dependency problem that motivates much research on the problem of estimating age, period and cohort effects is really the least significant aspect of the problem of estimating equation (5). The real problem is finding more and better proxies, better explanatory variables and sharper behavioral models that eliminate the vacuity inherent in context free statistical accounting schemes.

REFERENCES

Aigner, D. *et al.* (1984). Latent variables in econometrics, in Z. Griliches and M. Intrilligator (eds.), *Handbook of Econometrics.* North Holland, forthcoming.

Cagan, P. (1965). Measuring quality changes and the purchasing power of money: An explanatory study of automobiles. *National Banking Review* **3**, 217–236. Reprinted in Z. Griliches (ed.), *Price Indexes and Quality Change.* Cambridge, Mass.: Harvard University Press, 1971.

Easterlin, R. (1961). The American baby boom in historical perspective. *American Economic Review* **51**, 869–911.

Easterlin, R. (1978). What will 1984 be like? Socioeconomic implications of recent twists in the age structure. *Demography* **15**, 397–432.

Feller, W. (1968). *An Introduction to Probability Theory and Its Application.* New York: Wiley.

Fienberg, S.E. and Mason, W.M. (1978). Identification and estimation of age-period-cohort models in the analysis of discrete archival data, in K.F. Schuessler (ed.), *Sociological Methodology 1979.* San Francisco: Jossey-Bass, pp. 1–67.

Fienberg, S.E. and Mason, W.M. (1985). Specification and implementation of age, period and cohort models. This volume.

Goldberger, A.S. (1968). *Topics in Regression Analysis.* London: MacMillan.

Hall, R.E. (1971). The measurement of quality change from vintage price data, in Z. Griliches (ed.), *Price Indices and Quality Change.* Cambridge, Mass.: Harvard University Press.

Hanuschek, E. and Quigley, J. (1978). Implicit investment profiles and intertemporal adjustments in relative wages. *American Economic Review* **68**, 67–79.

Johnson, T. and Hebein, F. (1974). Investment in human capital and growth in personal income 1956–1966. *American Economic Review* **64** (4), 604–613.

Johnson, G. and Stafford, F. (1974). Lifetime earnings in a professional labor market: Academic economists. *Journal of Political Economy* **82**, 549–570.

Jöreskog, K. and Goldberger, A. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* **70**, 631–639.

Lazear, E. (1976). Age, experience and wage growth. *American Economic Review* **66** (4), 548–580.

Mincer, J. (1974). *Schooling, Experience and Earnings.* New York: Columbia University Press.

Mincer, J. (1978). Comments on the Hanuschek-Quigley paper. Unpublished mimeo, Columbia University.

Mincer, J. and Polachek, S. (1974). Family investments in human capital: Earnings of women. *Journal of Political Economy* **82**, March/April, S.76–S.109.

Nerlove, M. (1967). Distributed lags and unobserved components of economic time series, in W. Fellner (ed.), *Ten Economic Essays in the*

Tradition of Irving Fisher. New York: Wiley, pp. 120–169.

Weiss, Y. and Lillard, L. (1978). Experience, vintage and time effects in the growth of earnings: American scientists, 1960–1970. *Journal of Political Economy* **86** (3), 427–447.

Welch, F. (1979). The effects of cohort size on earnings: The baby boom baby's financial bust. *Journal of Political Economy* **87** (5), 565–599.

6. AGE-PERIOD-COHORT ANALYSIS AND THE STUDY OF DEATHS FROM PULMONARY TUBERCULOSIS*

William M. Mason

Herbert L. Smith[†]

Our purpose is both substantive and methodological. Substantively, we crystallize the results of prior research and expectations into an extended rationale for the application of the age-period-cohort accounting framework to the problem of understanding historical variability in the rate of tuberculosis mortality. This framework is then used to analyze a ninety year data series of tuberculosis mortality rates for the state of Massachusetts and a similar forty year series for the United States. The age-period-cohort accounting framework yields age effects with an expected pattern not well understood, period effects consistent with the advent of successful chemotherapeutic regimes after World War II, and steadily declining cohort effects whose interpretation has yet to be verified. In an attempt to pin down a possible interpretation, we show that cohort nativity composition affects the trend of cohort mortality in the Massachusetts series, and both level and trend in the United States series. Our findings consolidate the results and anticipations of past research on TB mortality based in part on two-effect models and graphic display of rates, and to some extent clarify various proposed interpretations of the historical trend; they leave open the ultimate

* We thank Carol Carter for invaluable help in obtaining the source data for Massachusetts, Albert Anderson and Jonathan Sell for programming assistance, Carol Crawford for typing, and David Freedman, Warren Sanderson and the other conference participants for their thoughtful comments on the first draft. We are especially indebted to Stephen Fienberg for statistical advice, Halliman Winsborough for encouragement and comments, and Otis Dudley Duncan, whose skeptical reaction to an earlier draft ultimately led us to discover a major error which this draft corrects.

† The authors' names are listed in alphabetical order; this is truly joint work. William M. Mason is affiliated with the Department of Sociology and the Population Studies Center, The University of Michigan, and Herbert L. Smith is affiliated with the Department of Sociology, Indiana University.