Identifying Age, Cohort and Period Effects in Scientific Research Productivity:

Discussion and Illustration Using Simulated and Actual Data on French Physicists

Bronwyn H. HALL (UC Berkeley and NBER) [bhhall@econ.berkeley.edu]
Jacques MAIRESSE (CREST and NBER) [mairesse@ensae.fr]
Laure TURNER (ENSAE and CREST) [laure.turner@ensae.fr]

26 June 2005

Abstract

Identifying Age, Cohort and Period Effects in Scientific Research Productivity:

Discussion and Illustration Using Simulated and Actual Data on French Physicists[1]

Bronwyn H. HALL, Jacques MAIRESSE, and Laure TURNER

1 Introduction

Empirical studies in the social sciences often rely on data and models where a number of individuals born at different dates are observed at several points in time, and interest centers on the identification of age, cohort, and time or period effects in the relationship of interest. However, modeling and identification of such relationships has proved to be problematic, largely because of the obvious impossibility of observing two individuals at the same point in time that have the same age but were born at different dates. The identification problem is further aggravated if one uses standard panel data estimators in which one takes first differences (or within individual differences) of the variables, in order to control for unobserved individual effects. In this case, the cohort effect disappears completely (because it is collinear with the individual effects), which obscures but does not eliminate the problem of identifying year and age effects simultaneously.

A number of "solutions" to this identification problem have been offered in the literature in different contexts (e.g., R. E. Hall 1971, Mason et al. 1973, Rodgers 1982a,b, Mason and Fienberg 1985, Berndt and Griliches 1991), all of which assume restrictions on the specification of the general underlying model, usually by imposing some sort of functional form assumption on the way the three effects enter. Hall (1971) was concerned with disentangling depreciation (the age effect), embodied technical change (the cohort effect), and disembodied technical change (the period effect) in a vintage capital model applied to trucks, in which he imposed the constraint that the two most recent vintages were identical in order to identify the model. Berndt and Griliches (1991) were interested in a problem similar to that

[1]This is a revision of a paper prepared for the SPRU Conference in Memory of Keith Pavitt at the University of Sussex, 13-15 November, 2003. We are extremely grateful to Serge Bauin and Michele Crance from UNIPS-CNRS, France for their invaluable help in constructing the database of condensed matter physicists.

confronting Hall: the construction of a hedonic pricing model of personal computers that incorporates technical change, vintage, and age effects. Unlike Hall, they explored and exposited the full range of assumptions available for identification of the additive dummy variable model.

At the same time, the problem had not gone unnoticed in the sociological literature, especially as it related to the interpretation of cohort effects. In a series of papers William Mason and his co-authors proposed estimating cohort-age-period models using identification assumptions similar to the one used by Hall (1971). This work culminated in a conference volume published in 1985 (Mason and Fienberg 1985) that provides an excellent overview of the state of the art and the views of sociologists, statisticians, and economists on the problems associated with this kind of modeling, both conceptual and methodological.

One of the many domains in which this identification problem is prevalent is the study of the scientific productivity of researchers, where we would like simultaneously to take account of differing productivity over time, as a function of age, and as a function of the vintage of the researcher. Scholars in the sociology of science, and more recently economists, have tried to measure the age-related productivity curve, and to purge it of effects due to the vintage of the researchers and the periods in which they are being observed. A major problem in such analysis is the need to take into account two major tendencies: the exogenous increase of publications with time and with cohort. Descriptive statistics on scientific publications suggest that they tend to increase over time more or less rapidly in many scientific fields, overall but also per researcher. A way to capture such time effects, as well as any general changes in the state of art and work environment is simply to introduce period (year) indicators in the model. In the same manner, it seems that younger cohorts tend to publish more than older ones when they were the same age, which may be related to the fact that there are increased incentives and competition for the younger generations, and/or they are more motivated and better trained, and/or that the cost of publishing is less (with the use of computers and internet, and growing numbers of journals, etc.). However, including cohort indicators (or for that matter individual effects) together with period indicators in the model introduces the aforementioned identification problem with the age variable.

In this paper we give an overview of the general identification problem of age, cohort, and period effects in a panel data regression model, of the estimation and interpretation difficulties

it raises, and propose what we think a practical approach to deal with them (second section); we illustrate these difficulties and the suggested solutions on simulated data (third section), and on a rich longitudinal database of the publications over 20 years (1980-2002) of about 500 French condensed matter physicists (fourth section). We have three goals in undertaking this work: 1) to illustrate the potential for such data to lead to misleading inference if the identification problem is overlooked or not confronted; 2) to discuss the estimation and interpretation of cohort-age-period models when there are individual effects; 3) to apply our methods to a panel of real data in order to draw some conclusions about the evolution of scientific research productivity over time and age.

We do not break new statistical ground on these questions here. Instead we outline how to apply the methods proposed by previous researchers to the problem of scientist productivity and we explore the implications of the resulting estimates for substantive research questions, in particular to highlight the ambiguous nature of some of the previous results in this area. To put it another way, we want to underline the importance of *a priori* assumptions in interpreting results from a cohort-age-period regression. Our research questions are the following: How do we interpret results when there is more than one way to achieve identification? What happens when we remove the individual effects (effectively removing the cohort information) and how do we interpret the results in this case?

2 The Age, Cohort and Period Identification Problem

2.1 Problem statement

It is well-known that the identity age = year (period) - year of birth (cohort) implies that all three effects cannot be identified in a linear model. It is somewhat less well-known that identification can be achieved in a dummy variable model by dropping a small number of variables (e.g., see Berndt and Griliches 1991). In fact, no experiment can be devised to identify a completely general model with cohort (C), age (A), and period (P) effects. Given the identity $A = P-C$ that exists in the data, it is obvious that any function $f(C, P, A)$ can be written as $f(C, P, P-C) = f(C,P)$, so that it does not depend on the value of A. It will therefore always be necessary to impose some constraints or prior information on $f(.)$ if we wish to

identify an age effect that is parsimonious and not simply derived from the cohort-period behavior.[2]

There exists a large body of prior research and debate in sociology, demography, and economics over the question of exactly how to identify all three effects using suitable constraints on the functional form of the relationship or other prior information. In sociology a rather heated debate over identification between William Mason and his co-authors and Willard Rodgers was conducted in the pages of the *American Sociological Review* in 1982 (Mason et al. 1973; Rodgers 1982a,b; Smith et al 1982). Mason et al. (1973) had proposed a method of identifying a model with three sets of dummy variables for age, period, and birth by constraining some of the coefficients, and Rodgers critiqued their approach strongly because of its ad hoc nature, arguing that a better method of identification was to replace one of the sets of dummy variables with 'real" variables that were correlated with that particular aspect of the relationship (i.e., replacing period dummies with variables describing the macro-economy during the period). Part of his critique was based on the argument that modeling the effects as additively separable already imposed too many constraints on the model and did not allow for interactions between, for example, cohort and changes over time.

Nevertheless, most researchers who are interested in identifying three separate effects begin by assuming that they are additive, that is, that

$$f(C,P,A) = f_C(C) + f_P(P) + f_A(A) \tag{1}$$

Clearly when the $\{f_J\ ,\ J = C,\ P,\ A\}$ are linear we have the well-known case that one of the three functions is not identified. However, Heckman and Robb (1985) show more generally that when the f_J are polynomials of order J, only $\begin{pmatrix} J+1 \\ J \end{pmatrix}$ combinations of the $\begin{pmatrix} J+2 \\ J \end{pmatrix}$ coefficients on the terms of order J are identified. That is, for the linear model, only 2 of the 3 linear coefficients are identified. For a quadratic model, only 3 of the 6 quadratic coefficients

[2] The requirement for parsimony is another way of saying that we expect the age effect to be rather smooth and slow to change and that we would like to impose that belief on the model. Conceptually if it were not for our *a priori* belief that things change slowly with age, we could simply derive the age effect from the observed cohort and period effects via the identity. In fact, some scholars (notably Rodgers) would argue that this is all that can be done in any case without external prior information such as the macroeconomic environment. That is, the identification problem is fundamental, given the impossibility of observing A such that $P\text{-}C \neq A$.

are identified, and so forth. So although low-order polynomials seem an attractive way to model these effects because of their smoothness, in practice they have not been much used because the lack of identification is so obvious.

Given additivity, the most general semi-parametric model is a model that simply includes dummy variables for all three effects. However, if we do not impose additivity, a more general model is available, one which is simply the means of the dependent variable for each cohort-period combination. If there are no covariates other than cohort, age, and period, these means are the sufficient statistics for the data.[3] In the next section of the paper we begin with this model as our baseline and then present a series of models that are nested within it.

2.2 Model with age, cohort, and year dummies

Suppose that we have data on a variable of interest Y_{it} on N individuals from C cohorts, observed for P periods. If we have no prior information on the relationship of Y to cohort and period other than assuming that it is multiplicative in levels (and therefore additive in logarithms), the natural semiparametric regression model simply includes a dummy variable for each cohort-period combination. Such a method uses all the information available from the means of the data by cohort and period (and therefore age), and exhausts the degrees of freedom.

Using lower case y to denote the logarithm of Y, this model can be written as

$$\text{saturated:} \quad y_{it} = a_{ct} + \varepsilon_{it} \tag{2}$$

where $i = 1,\ldots,N$ individuals; t = $1,\ldots,P$ periods; and $c = 1,\ldots,C$ cohorts. We are implicitly assuming that the data are balanced across P and C (although not necessarily across N). That is, for each cohort we observe a complete set of P periods.[4] Given the assumption of balance in the P and C dimension, when we observe PxC cells, we are observing $A = P+C-1$ ages.

[3] Strictly speaking, we would also need the variance of the dependent variable and an assumption of normality and conditional homoskedasticity for these means to be sufficient.

[4] Symmetrical treatments where the data are balanced for C and A (we observe the same number of ages for each cohort), or where the data are balanced for P and A (we observe the same number of ages in each period, and therefore cohorts are unbalanced) are possible. We present the C and P case here because that is the way our data is organized: the changes necessary to estimate with data balanced for CA or PA are obvious but tedious.

This model, which we call the saturated model, allows us to identify *PC* means of *y*, one for each cohort-period combination. However writing the model this way does not provide estimates of age, cohort, or period effects separately, nor does it impose constancy on these effects.

As the saturated model is the most general model that can be estimated using this type of data, it is a useful starting point, but most researchers prefer to impose constancy of the coefficients across the same ages, cohorts, and periods, which leads to a model that we call the *threeway* or *CAP* model:

$$\text{CAP:} \quad y_{it} = \mu + \alpha_c + \beta_t + \gamma_a + \varepsilon_{it} \tag{3}$$

We know that one cannot estimate equation (3) directly: the coefficients of the different indicators can only be estimated relative to a reference value for each of the three dimensions. Therefore one imposes (for example) nullity on the coefficients α_1, β_1, and γ_1, which are respectively those of the first cohort, the first period, and the first age. However, the collinearity between the indicators of age, period, and cohort has not been removed by this procedure: in fact, it is easy to show that even the variables in this new equation will not be linearly independent. How can one then estimate this model? Several methods have been proposed in the literature and their use was the subject of considerable debate for a time.

Mason *et al.* (1973) proposed determining the number of restrictions which it is necessary to impose on equation (3) in order to eliminate the problem of collinearity and identify the model. They demonstrated that one possible sufficient condition is to constrain two coefficients in the same dimension (age, period, or cohort) to be equal. For example, by imposing that the effects of the first and last ages are equal, one can identify the model, provided that there are at least 12 cohort-period combinations. The number of coefficients that can be estimated is therefore $1 + (P-1) + (C-1) + (A-1) - 1 = 2(P+C) - 4$, as compared to PC for the saturated model. When $P=C=2$, the three-way model coincides with the saturated model, implying that at least one of P or C must be larger than 2 for this model to impose meaningful constraints.

Naturally, the problem with identifying the three-way model using an equality constraint on two of the coefficients is that the different equality constraints will correspond to different estimates of the coefficients. The explanatory power of the models (measured by the R-squared) estimated under the different equality constraints will be the same. As a consequence, in the absence of an equality constraint that is preferred *a priori*, identifying the model in this way does not allow the selection of a "good" model. A secondary problem is that the identification may be fairly weak, relying as it does on a single equality constraint between coefficients.

However, as Berndt et al. (1995) showed, when the number of periods and cohorts is large enough, the three-way model imposes a number of constraints on the saturated model that can be tested in order to determine its plausibility. Similarly, models with only two or one set of dummies are nested within the three-way model, so that it is possible to test their validity using either the saturated or the threeway model as the maintained hypothesis. We write the two-way models as follows:

$$\begin{aligned} \text{CP:}\quad & y_{it} = \mu + \alpha_c + \beta_t + \varepsilon_{it} \\ \text{CA:}\quad & y_{it} = \mu + \alpha_c + \gamma_a + \varepsilon_{it} \\ \text{PA:}\quad & y_{it} = \mu + \beta_t + \gamma_a + \varepsilon_{it} \end{aligned} \tag{4}$$

and the one-way models similarly:

$$\begin{aligned} \text{C:}\quad & y_{it} = \mu + \alpha_c + \varepsilon_{it} \\ \text{P:}\quad & y_{it} = \mu + \beta_t + \varepsilon_{it} \\ \text{A:}\quad & y_{it} = \mu + \gamma_a + \varepsilon_{it} \end{aligned} \tag{5}$$

For example, testing the *CP* model against the *CAP* model is equivalent to testing whether the *A-1* coefficients $\gamma_2, \gamma_3, \ldots, \gamma_A$ are equal to zero, which corresponds to testing the constraints on the saturated model given in Table 1.

[Table 1 about here]

Similar tables apply for the other models.[5] The implication of this particular set of constraints (those for the *CP* model) is that the change in *y* from period to period is the same for each cohort, but that the change in *y* from age to age is different for each cohort. The number of constraints relative to the saturated model is equal to *PC – (P-1) – (C-1) – 1 = (P-1) (C-1)*, which can be a sizable number. In section 4 of the paper we present empirical results for a panel of French physicists which has *25* cohorts and either *12* or *21* periods. For the shorter sample using these data, the number of implied constraints for the *CP* model is equal to *264* out of *300* coefficients. Table 2 gives the general formulas for the number of constraints in all the models when the data are balanced in the cohort and period dimension. Table 3 illustrates the computations for our two panels, where we have $i = 1,\ldots,N$ individuals ($N=418$); $p = 1,\ldots,P$ years ($P=12$ or 21); $c = 1,\ldots,C$ cohorts ($C=25$); and therefore age $a=t-c$ ($A=36$ or 45).

[Tables 2 and 3 about here]

It is clear from these tables that when there are a large number of years or cohorts, there are a large number of implied constraints. The implication is that even though it is not possible to identify a model with a full set of cohort, year, and age dummies, it is still possible to test for the presence of any one set of these dummies conditional on including the other two sets. That is, because only one additional constraint is required to identify the model with all three effects, when more than one additional constraint is implied by dropping a set of dummies, we can still perform a test. As mentioned earlier, in the case of data balanced in the cohort and period dimension, this will be true when either the number of periods or the number of cohorts is at least three.

2.3 Including individual effects

In many situations, it is desirable to control not only for effects due to the cohort to which an individual belongs, but for permanent differences in individuals as well, leading to a variation of the CAP model:

$$\text{IAP:} \quad y_{it} = \mu_i + \alpha_c + \beta_t + \gamma_a + \varepsilon_{it} \tag{6}$$

It is obvious that this will create a further identification problem: given any individual *i,* the cohort *c* to which he belongs is known, and the cohort effect α_c is therefore completely

[5] An appendix gives the details for the *CA* and *PA* models.

unidentified in a model with individual effects. In addition, some of the identification strategies discussed above (specifically those involving constraining the cohort dummies) are unavailable, because including individual effects necessarily involves including a complete set of cohort dummies. One additional danger with including individual effects in this models (and as a consequence differencing out the cohort effect) is that the identification problem itself is therefore obscured and may be missed by the researcher.

Heckman and Robb (1985) discuss the identification issue in CAP models with individual effects and suggest an alternative identification strategy using a variance components decomposition. That is, they propose modelling using random effects in cohort, age, and period, and then estimating the model using the moment matching methods associated with Joreskog's LISREL program.

3 An illustration using simulated data

In order to illustrate the identification problem and the difficulties it creates for measuring age effects in researcher productivity, we performed a series of simulations using data calibrated to match the panel of French physicists analyzed in Turner and Mairesse (2005) and also in section 4 of this paper. That dataset had observations on the publications of 465 individuals who were born between 1936 and 1960 (25 cohorts), for the period 1986 to 1997 (12 years). In this section of the paper we show the results of a typical simulation, first graphically, and then as a series of statistical tests designed to choose the correct model. The particular simulation we chose illustrates the potential for a model that has only cohort and period effects to generate data that may appear to have peak in productivity at a certain age in spite of there being no age effect in reality.

Our approach here is to generate data that looks like the real data using a negative binomial model (so we obtain counts with overdispersion), but to estimate using the log-linear dummy variable model that is common in the literature. Given the generally small values of the dependent variable and the fact that we are using dummy variables, the differences between using OLS or using the more correct ML on a negative binomial model for estimation are likely to be slight. Figures 1a-1c show the results of simulating the model given below:

$$y_{it} \sim NB(\lambda_{it}, \sigma^2)$$
$$\lambda_{it} = \mu_0 + \alpha t + \beta c + \gamma c^2 \quad (7)$$

where *NB* denotes the negative binomial distribution, t is the period (1986-1997), c is the cohort (1936-1960), $\alpha = .05$, $\beta = 20$, $\gamma = -.00513$, and μ_0 and b were chosen so that the simulated data had the same mean and variance as the actual data, whose mean was 2.7 publications per year with a standard deviation (conditional on cohort and period) of 3.2. These parameter values imply that the quadratic in c reaches its maximum in about 1949, in the middle of our data period, but that the slope never exceeds about 0.15 publications per year in absolute value for the observed cohorts and is usually much lower, of the same order of magnitude as the year effect (*.05* publications per year).

[Figures 1a-1c about here]

Each panel of Figure 1 shows the resulting data from this simulation, plotted three different ways: 1a shows the means by age, 1b the means by year, and 1c the means by cohort. In each case we also show the best fit line for the dummy variable model that excludes the variable on the *X* axis, as a guide to the eye. Note that any dummy variable model which includes a set of dummies for the *X*-axis variable will fit the means of the data perfectly. For example, in Figure 1a we show the fit from a model that includes only the cohort and period dummies (the CP model). Any model that includes age dummies (that is, the CAP, CA, PA, or A models of Section 2) would have matched the overall age means exactly. Of course, were we to examine the fit of the age distribution for particular cohorts or particular years, only the saturated model would be able to match the data exactly. This fact is illustrated in Figure 2, which shows the data and the fit of the various models for three separate cohorts (1936, 1948, and 1960) that have three sets of non-overlapping ages (50-61, 38-49, and 26-37).

[Figure 2 about here]

The main message of Figure 1 is that although the year and cohort distributions look the way we would expect, given the simulation, the resulting age distribution exhibits smooth behavior with peaking during the 40s, even though there is no age effect in our simulated model. As we expected, the year distribution shows a modest trend increase of about 0.6 publications on average throughout the twelve-year period and the cohort distribution a slight peaking

tendency in the late 1940s. Our conclusion is that for samples of our size, averaging approximately 17 observations per period-cohort cell, it would be possible to observe a peaked age effect even if one is not there, at least if there is curvature in the cohort or period dimension.[6] That is, the observed age effect can be generated simply by the interaction of period and cohort effects.

What are the implications of this "age" effect for model selection? That is, even though we observe something that looks like an age effect in Figure 1a, the testing strategy outlined in section 2 of the paper may allow us to choose correctly among the many possible models that are given in Tables 2 and 3, at least when the number of cells or observations are large enough, and to reject models that are inappropriate for the data. The tests corresponding to the eight different models in Table 2 are nested in the way shown in Figure 3: the threeway CAP model is nested within the saturated model, the three twoway CP, CA, and PA models are nested within the CAP model, and the three oneway C, P, and A models are nested within either of their corresponding twoway models. Thus we can test for the correct specification using a general-to-specific sequence of tests: first we test the CAP model using the saturated model as the maintained hypothesis, and if we accept, then we can test the three twoway models using the CAP model as the maintained hypothesis, and so forth.

[Figure 3 about here]

For the simulated data shown in Figures 1 and 2, the results of this model selection approach are somewhat ambiguous. Given the saturated model, we can easily accept a model with cohort, year, and age effects only, but conditional on that model, there are two models that will describe the data accurately: one is the cohort-year model, which is consistent with the data generating process, and the second is the cohort-age model, which is not. The year-age model and the three oneway models are clearly rejected, regardless of the model that is taken as the maintained hypothesis. Our conclusion is that for data like ours, it may be difficult to discriminate between some of the models using samples of the size available to us, although clearly we are able to reject the more restrictive specifications.[7] In the next section of the

[6] In this investigation we have focused on a quadratic age effect because that is a typical finding of the human capital literature and is therefore of considerable interest to researchers of scientific productivity. Spurious linear age effects would be even easier to generate using trends in period and cohort.

[7] Discuss the lack of true simulation results.

paper we apply the same sequence of tests, this time to the real data, and reach similar conclusions.

4 An application using data for a panel of French physicists

There are many studies of age and/or gender differences in research production in the sociology of science and in scientometrics (for example Cole, 1979; Cole and Zuckerman, 1984; Cole and Singer, 1991; see also Stephan, 1996). Economists have also investigated them in the framework of cumulative advantage models and/or life cycle models (Diamond, 1984; Levin and Stephan, 1991; David, 1994; Stephan, 1998). These models reveal the consequences of events arriving in the early career of the scientist on the one hand and of the anticipation of the coming end of career on the other on the allocation of research efforts over time and individual productivity. However, there has been relatively little research based on individual panel data, which could allow disentangling the effects of age and gender from cohort and period effects, as well as from other unobservable individual effects. One of the few exceptions is Levin and Stephan (1991), in which the proposition that research activity declines over the life cycle is tested on publication panel data for scientists in six sub-fields of earth science and physics (including condensed matter physics), over the period 1973-1979.

4.1 The dataset

The database with which we work is an original panel database that was created from the records of 523 French condensed matter physicists working at the CNRS between 1980 and 2002, and born between 1936 and 1960. Condensed matter (solid state) physics comprises half of all French academic physics. During the period of study, it was a rapidly growing field with relatively little mobility towards the private sector or the universities, and with well-identified journals.[8] The group of physicists studied here represents a majority of all CNRS researchers in this discipline (they numbered 598 in 2002). The CNRS and universities are the main public research institutions in this domain in France. In 2002, 28.3% of the condensed matter physicists belonged to the CNRS and 70.5% to the academic sector (1489 researchers).

[8] For further information on the database and its creation, see Turner (2003).

Our panel database is unbalanced both because the scientists enter at different dates, and because some exit before 2002.[9] We restrict the analysis in this paper to a panel analyzed by Turner and Mairesse (2005) containing 465 physicists, observed from 1986 to 1997, aged 26 to 60 and with twelve years of data. Tables 4 and 5 contain some simple statistics for our data. 18 per cent of these researchers are female, rising from 15 per cent in the earliest cohort (those born 1936-40) to over 20 per cent in the last two cohorts (those born 1951-1960). About the same number have a doctorate degree from a Grande Ecole, of this number about 16 per cent are female.[10] Over half of them (62 per cent) started their career in either Grenoble or Paris, which are considered the most important centers in this field. Almost half of the researchers changed labs at least once during their career. The median number of researchers in the labs in which they worked was 43, and the sample published at a slightly higher rate than their labs (2.7 papers per year versus 2.3 for the average researcher in the lab).

As is usual in this literature (Levin and Stephan 1991), our measure of researcher productivity is the count of articles published during the year.[11] The total number of articles published is about 15,010 (2.7 per person per year) but 25% of the observations have no publications in a given year and one individual has 62. Fitting a simple Pareto distribution to these data yields a coefficient of about 0.1, which implies that the distribution has neither a mean nor a variance. Figures 4 and 5 show the smoothed sample averages of the productivity measure plotted versus age and calendar year respectively, for five year groupings of the cohorts (year of birth). As expected, the average number of articles published tends to increase over time, although the main differences seem to be by cohort rather than year, with the exception of the most recent cohort. The age distributions for the earlier cohorts suggest a peak somewhere in the late 40s or early 50s, although not very strongly.

---

[9] The identification problem is complicated in our setting by the fact that there is a small amount of variation in the identity given above due to entry at different ages (90% of the researchers enter between age 23 and 32) which yields apparent identification, but where such identification is achieved using only a few of the observations. In this paper we abstract from this complication by defining age to be year less entry cohort rather than calendar age.

[10] Explain the significance of the meaning of a Grande Ecole degree.

[11] We also have several other measures available: articles published weighted by the number of co-authors, the average number of pages in an article, and measures based on citations received in the first two and five years, weighted by the impact factors for the journals in which the citing papers appeared (the average citation rate of its articles). However, in the present paper we focus on the article count itself, which is suficient to illustrate the various identification strategies. See Table 5 for simple statistics on the other measures.

[Tables 4 and 5 about here]

[Figures 4 and 5 about here]

4.2 Productivity and age

In this section of the paper we use the tests described earlier to ask whether the apparent peak in productivity as a function of age can be due to the confluence of cohort and year effects. In Figure 6 we show the results of our tests applied to the actual data on French solid state physicists. The results are similar to those for the simulated data. The preferred specifications with only two sets of dummies are those with cohort and year or cohort and age effects, although they are both rejected at the 5 per cent level in favor of a specification with all three sets of dummies. Nevertheless, note that the test for a model with only cohort and year effects versus that which includes age effects in addition has a p-value of 0.034, which is fairly large given the number of observations (5580). The conclusion is that the independent effect of researcher age above and beyond that due to the cohort in which he or she entered and the year of publication is slight.

[Figure 6 about here]

Alternatively, if we prefer a specification with cohort and age effects only, that is a model where calendar time influences only the "initial condition" for the researcher, such a model would be only marginally less preferred to the cohort-year model. The point is that in order to distinguish these alternatives it will be necessary to appeal to some prior information, as the data themselves cannot really tell us which is correct.

To underline this point, we show the actual and fitted values from the two models in Figures 7 and 8, plotted first versus age and then versus time. Figure 7 shows the geometric means of the data (publication counts) for each age, and the geometric means of the values predicted by the cohort-year model (those predicted by the cohort-age model will lie precisely on the actual data).[12] Similarly, Figure 8 shows the same thing by time, with the fitted values from the cohort-age model, since the cohort-year predictions will coincide exactly with the data when it

[12] Geometric means are used because the model was fit in log-linear form, so these are the unbiased predictions (but without the correction for the residual variance, which is small).

is displayed in this way. Looked at in the age dimension, the cohort-year model appears to miss a bit at the youngest and oldest ages, although it does reproduce the slight peaking. Looked at in the time dimension, the cohort-age model appears to impose an acceptable smoothness on the data. So from this perspective we might prefer the cohort-age model, even though the fit of the two models is nearly identical.

Now suppose the research question concerns the age at which publication productivity peaks. In this case the choice of model may matter. For example, consider the choice between the threeway model and the cohort-age model, both of which will reproduce the data means when looked at in the cohort-age dimension. Nevertheless, the two models may predict a different productivity peak. A quadratic fit to the two sets of age dummies obtained from these two models using our data yielded the following result: research productivity peaks at 52.2 years of age using the threeway model and at 53.7 years of age using the cohort-age model and ignoring the calendar time effects if they are there. Although this difference is not large, it is significant.[13]

[Figures 7 and 8 about here]

But that is not the end of the problem. Consider the following model which combines a quadratic in age with a set of year dummies and and a set of cohort dummies:

$$y_{it} = \mu + \alpha_c + \beta_t + \gamma_1 a_{it} + \gamma_2 a_{it}^2 + \varepsilon_{it} \tag{8}$$

At first glance, this model looks sensible and in fact has often been estimated, sometimes with individual effects rather than cohort effects included (Levin and Stephan 1991; Turner and Mairesse 2005). Identification (with an intercept included) requires omission both of one of the cohort dummies and of one of the year dummies. However, because age (a_{it}) is an exact linear function of cohort and period, which identifying assumption you choose (and there are an infinite number) will affect the estimates of γ_1 and γ_2, and therefore, the estimate of the age at which productivity peaks (which is - $\gamma_1/2\ \gamma_2$).

[13] If a quadratic model in age is included directly in the model with cohort-year dummies and that with cohort dummies alone, the difference in peak age is even larger: 50.6 years versus 53.8 years.

Figure 9 shows a representative result for our data. The identifying assumptions used were to include a complete set of year dummies, exclude the intercept, and include all but one of the cohort dummies. The excluded cohort dummy was allowed to vary from 1936 to 1960. The figure shows a few representative examples of the resulting age profile (excluding year and cohort effects). Note that all the fits were identical, in the sense that the sum of squared residuals were exactly equal, and they all generated the same age-cohort-year means, but very different age-productivity profiles. The problem is interpretive: the age-cohort-year identity means that it is impossible to identify the productivity curve as a function of age without strong prior restrictions on the year and cohort effects (such as their absence). The age at which productivity peaks also varies significantly for the different normalizations: it is 39.6, 41.3, 42.4, 37.9, and 50.4 when we drop the dummy for entry in 1936, 1940, 1944, 1948, and 1952 respectively.

5 Conclusions

This paper has explored a familiar identification problem, that of vintage, age, and time, in a context where it does not seem to have been sufficiently recognized: the identification of cohort, age, and period effects in scientific productivity. We have emphasized the fact that identifying an age-related productivity effect or the presence and location of a productivity peak relies crucially on what we are willing to assume about the variation in the other two dimensions, cohort and time. There is no universal solution to this problem, given the identity that relates the three.

Therefore we recommend the following: test for the presence of each of the threee effects semi-parametrically as we have done in this pape. If the tests reveal that one dimension can be ignored, then the most parsimonious specification will include only the other two dimensions. However, the power of such a test clearly goes up with the dimensions of the data: in some unreported experiments, we found that 12 years and 36 ages led to confusion between a cohort-age and a cohort-year model when the former was the true model, whereas 25 years and 44 ages allowed us to distinguish the two.

Alternatively, we return to the original recommendations of Rodgers, who strongly advocated the use of *a priori* information about cohorts or the time period to help identify the model. We

note that this approach was the one taken by Stephan and Levin (1991) when they achieved identification by grouping the cohorts in their sample according to the knowledge base to which they were exposed in their graduate training.[14] The method used by Turner and Mairesse (2005), grouping in five year intervals, seems somewhat less satisfactory in this context. This amounts to achieving identification of the age effect by comparing closely adjacent ages and assuming they come from the same cohort. In this case, it would seem preferable to use the actual variation in year of entry into the sample (cohort) for individuals of the same age rather than creating spurious age variation by holding the cohort fixed.

[14] However as we discuss in the paper, and as is clear from their detailed discussion of the tables in Levin and Stephan, this method of identification breaks down if individual rather than cohort effects are included.

References

- ARORA A., P. A. DAVID and A. GAMBARDELLA. 1999. "Reputation and Competence in Publicly Funded Science: Estimating the Effects on Research Group Productivity," *Annales d'Economie et de Statistique* 49/50: 163-198.
- BERNDT E. R., Z. GRILICHES and N. RAPPAPORT. 1995. "Econometric Estimates of Prices Indexes for Personal Computers in the 1990s," *Journal of Econometrics* 68: 243-268.
- BERNDT E. R. and Z. GRILICHES. 1991. "Price Indices for Microcomputers: An Exploratory Study, " In *Price Measurements and their Uses*, Chicago: University of Chicago Press, pp. 63-93.
- BLUNDELL R., R. GRIFFITH and F. WINDMEIJER, 1995, "Individual Effects and Dynamics in Count Data", discussion paper 95-03, department of economics, University College, London
- BONACCORSI A. and C. DARAIO. 2003. "A robust nonparametric approach to the analysis of scientific productivity," *Research Evaluation* 12 (1): 47-69.
- BONACCORSI A. and C. DARAIO. 2003. "Age effects in scientific productivity: The case of the Italian National Research Council (CNR)," *Scientometrics* 58 (1): 49-90.
- COLE S., 1979, "Age and Scientific Performance", *American Journal of Sociology*, 84(4): 958-977.
- COLE J. and H. ZUCKERMAN, 1984, "The Productivity Puzzle: Persistence and Change in Patterns of publications of Men and Women Scientists", in *Advances in Motivation and Achievement*, vol. 2, pp. 217-258.
- COLE J. and B. SINGER, 1991, "A Theory of Limited Differences: Explaining the Productivity Puzzle in Science", in *The Outer Circle: Women in the Scientific Community*, H. Zukerman, J. Cole and J. Bruer eds., Norton, New York.
- DAVID P., 1994, "Positive Feedbacks and Research Productivity in Science: Reopening Another Black Box", in *The Economics of Technology,* Ove Granstrand ed., Amsterdam: Elsevier Science: 65-89.
- DIAMOND A., 1984, "An economic model of the life-cycle research productivity of scientists", *Scientometrics*, 6(3): 189-96.
- HALL R. E., 1971, "The Measurement of Quality Change from Vintage Price Data," chapter 8 in Zvi Griliches (ed.) Price Indexes and Quality Change, Cambridge, MA: Harvard University Press, 240-271.

- HALL R. E., 1968, "Technical Change and Capital from the Point of View of the Dual," *Review of Economic Studies* 35: 35-46.
- HAUSMAN J. A., B. H. HALL and Z. GRILICHES, 1984, "Econometric Models of Count Data with an Application to the Patents-R&D Relationship", *Econometrica*, 52(4): 909-38.
- HECKMAN, J., and R. ROBB, 1985, "Using Longitudinal Data to Estimate Age, Period, and Cohort Effects in Earnings Equations," in W. Mason and S. Fienberg (eds), *Cohort Analysis in Social Research: Beyond the Identification Problem*, New York: Springer Verlag.
- LANCIANO-MORANDAT, C. and H. NOHARA, 2002, "The new production of young scientists (PhDs): a labour market analysis in international perspective," DRUID Working Paper No. 03-04.
- LEVIN S. and P. STEPHAN, 1991, "Research productivity over the life cycle: evidence for academic scientists," *American Economic Review*, 81(1):114-32.
- MASON K. O., W. M. MASON, H. H. WINSBOROUGH and W. K. POOLE, 1973 "Some Methodological Issues in Cohort Analysis of Archival Data," *American Sociological Review* 38(2): 242–258.
- MASON W. and S. FIENBERG (eds), 1985, *Cohort Analysis in Social Research: Beyond the Identification Problem*, New York: Springer Verlag.
- RODGERS W. L., 1982a, "Estimable Functions of Age, Period, and Cohort Effects," *American Sociological Review* 47(6): 774–787.
- RODGERS W. L., 1982b, "Reply to Comment by Smith, Mason and Fienberg," *American Sociological Review* 47(6): 787–793.
- SMITH H. L., W. M. MASON and S. E. FIENBERG, 1982, "Estimable Functions of Age, Period, and Cohort Effects: More Chimeras of the Age-Period-Cohort Accounting Framework: Comment on Rodgers," *American Sociological Review* 47(6): 787–793.
- STEPHAN P. E. 1996. "The Economics of Science," *Journal of Economic Literature* XXXIV: 1199-1235.
- STEPHAN P. E. 1998, "Gender Differences in the Rewards to Publishing in Academic Science in the 1970's", *Sex Roles*, 38(11/12).
- TURNER L. and J. MAIRESSE. 2005. "Individual Productivity Differences in Scientific Research: An Econometric Study of the Publication of French Physicists,"

Paper presented at the Zvi Griliches Memorial Conference, Paris, August 2003; Annales d'Economie et de Statistique, forthcoming.

- TURNER L., 2003, "La recherche publique dans la production de connaissances, contributions en économie de la science", PhD Dissertation, Université Paris 1 (http://www.crest.fr/pageperso/lei/laure.turner/these.htm)

Table 1

Constraints for the Cohort-Period Model

Periods/ Cohorts	**P_1**	**P_2**	**P_3**	**P_4**	**……**
C_1	$a_{11} = \mu$	$a_{12} = \mu+\beta_1$	$a_{13} = \mu+\beta_2$	$a_{14} = \mu+\beta_3$	…
C_2	$a_{21} = \mu+\alpha_1$	$a_{22} = \mu+\alpha_1+\beta_1$	$a_{23} = \mu+\alpha_1+\beta_2$	$a_{24} = \mu+\alpha_1+\beta_3$	…
C_3	$a_{31} = \mu+\alpha_2$	$a_{32} = \mu+\alpha_2+\beta_1$	$a_{33} = \mu+\alpha_2+\beta_2$	$A_{34} = \mu+\alpha_2+\beta_3$	…
….	…	…	…	…	…

Table 2

Number of Parameters and Constraints for the Different Models

Model	**Free parameters**	**Number of constraints**	**Minimum P, C for over identification**
Saturated	*P·C*	*0*	*NA*
CAP - cohort, age, and period	*P+C+A-3 = 2(C+P)-4*	*(P-2)(C-2)*	*P=3,C=3*
CP – cohort and period	*P+C-1*	*(P-1)(C-1)*	*P=2,C=2*
CA – cohort and age	*C+A-1 = 2C+P-2*	*(C-1)(P-2)*	*P=3,C=2*
PA – period and age	*P+A-1 = C+2P-2*	*(P-1)(C-2)*	*P=2,C=3*
C – cohort	*C*	*C(P-1)*	*P=2,C=1*
P – period	*P*	*P(C-1)*	*P=1,C=2*
A - age	*A = P+C-1*	*(P-1)(C-1)*	*P=2,C=2*

Table 3

Number of Parameters and Constraints for the Data

	Short Sample		**Long Sample**	
Model	**Free parameters**	**Number of constraints**	**Free parameters**	**Number of constraints**
Saturated	*300*	*0*	*525*	*0*
CAP - cohort, age, and period	*70*	*230*	*88*	*437*
CP – cohort and period	*36*	*264*	*45*	*480*
CA – cohort and age	*60*	*239*	*69*	*456*
PA – period and age	*47*	*252*	*65*	*460*
C – cohort	*25*	*275*	*25*	*500*
P – period	*12*	*288*	*21*	*504*
A - age	*36*	*264*	*45*	*480*

Table 4
Sample Statistics for 465 CNRS Physicists

Dummy variables

Description	Number	Share
Gender (1 = female)	84	18%
D (started in Grenoble)	121	26%
D (started in Paris)	167	36%
D (PhD from a Grande Ecole)	79	17%
Changed labs one or more times	205	44%

Table 5
Sample Statistics for 465 CNRS Physicists

Variables constant over time

Description	Median	Mean	Std. Dev.	Min	Max
Date of birth	1945	1946.8	7.3	1936	1960
Average lab productivity*	2.29	2.37	0.88	0.11	7.59
Average lab impact factor*	3.58	3.53	0.64	1.61	7.62
Intl openness - share art pub intl	0.037	0.039	0.028	0.000	0.109
No. of researchers in lab	43	46.4	26.3	0	134
No. of labs in career	1	1.61	0.79	1	4
Time-varying variables (5,580 observations for 1986-1997)					
Age of researcher this year	45	44.6	8.0	20	61
No of articles published in year	2	2.69	3.21	0	62
No of articles weighted by authors	0.20	0.21	0.19	0	1
Average number of pages	5.40	5.49	4.68	0	58
Impact factor (2 years)	2.54	2.66	2.30	0	21.48
Impact factor (5 years)	4.36	4.15	3.18	0	26.56

*Based on only 447 observations.

Figure 1a
Data Simulation with Cohort and Period Effects Only, Plotted versus Age

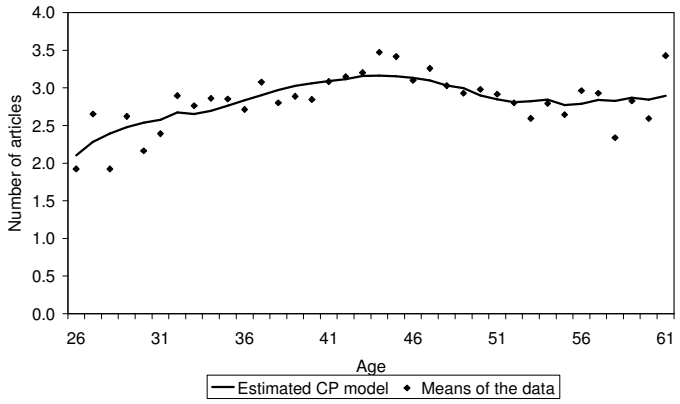


Figure 1b
Data Simulation with Cohort and Period Effects Only, Plotted versus Period

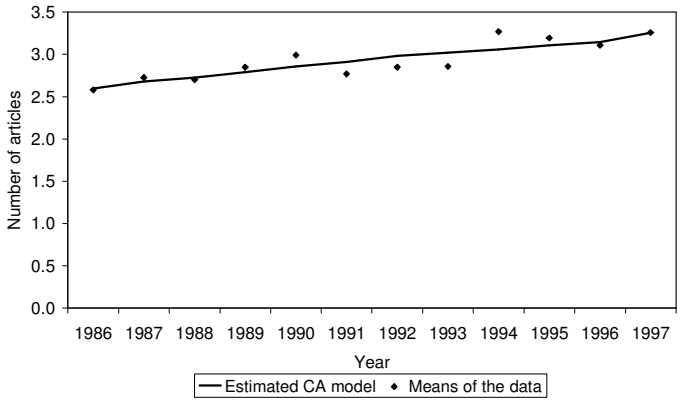


Figure 1c
Data Simulation with Cohort and Period Effects Only, Plotted versus Cohort

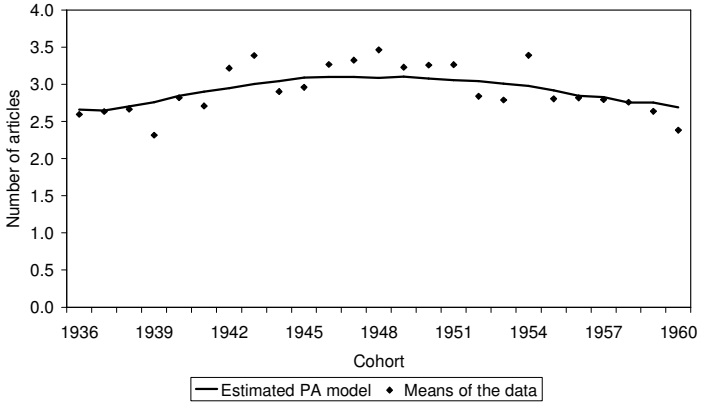


Figure 2

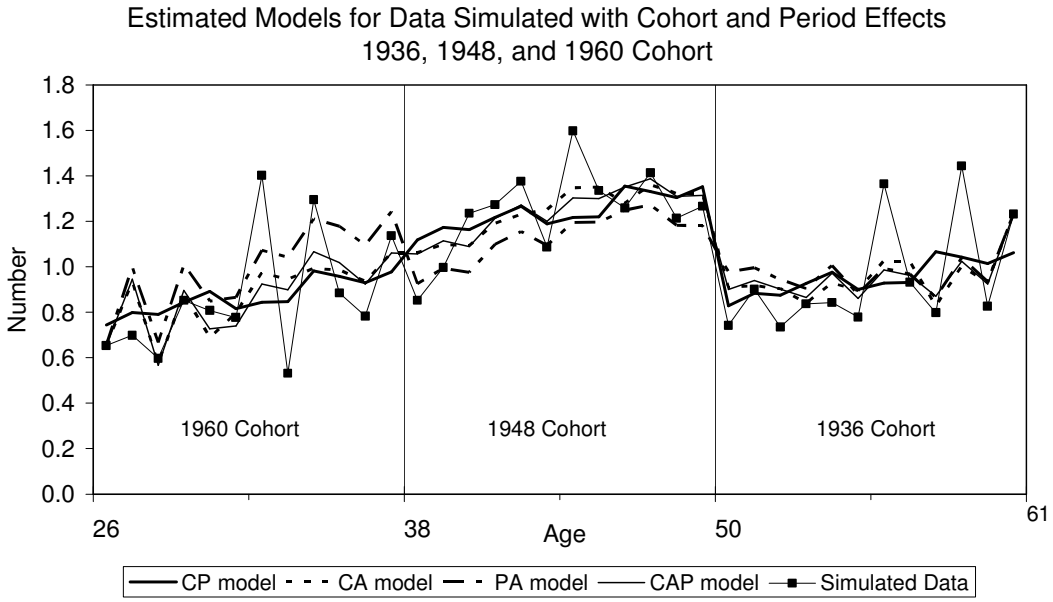


Figure 3
F-tests for Cohort, Age, and Period Models
Simulated Data

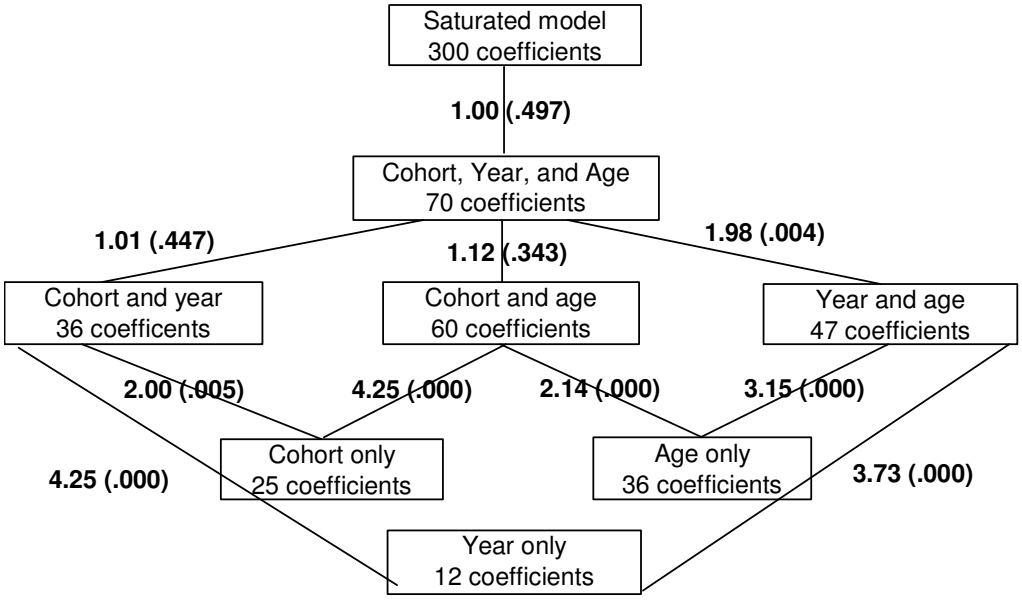


Figure 4
Average Number of Articles Published by Age
(5-year Moving Average)

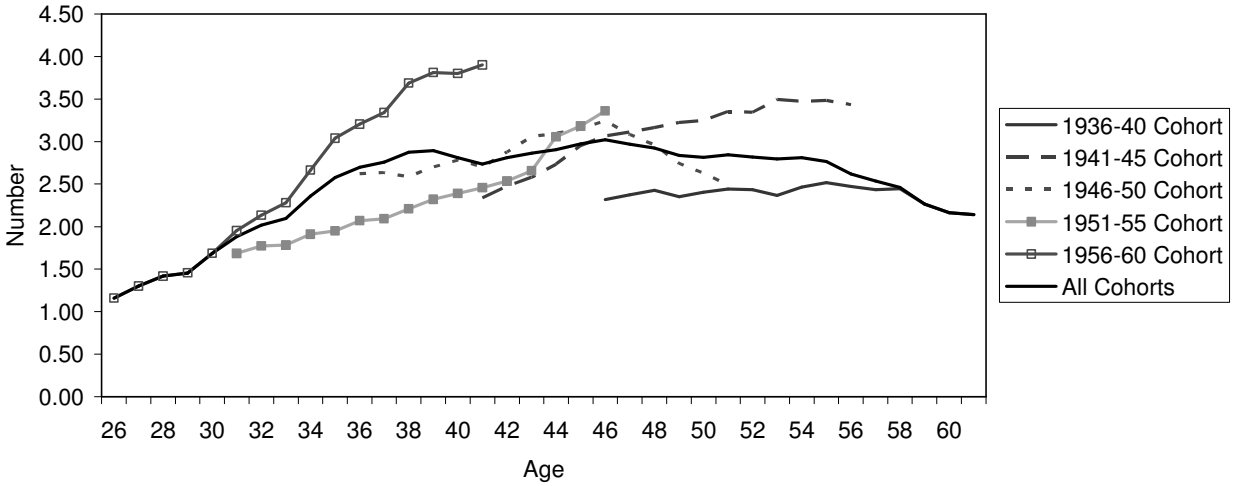


Figure 5
Average Number of Articles Published by Year
(5-year Moving Average)

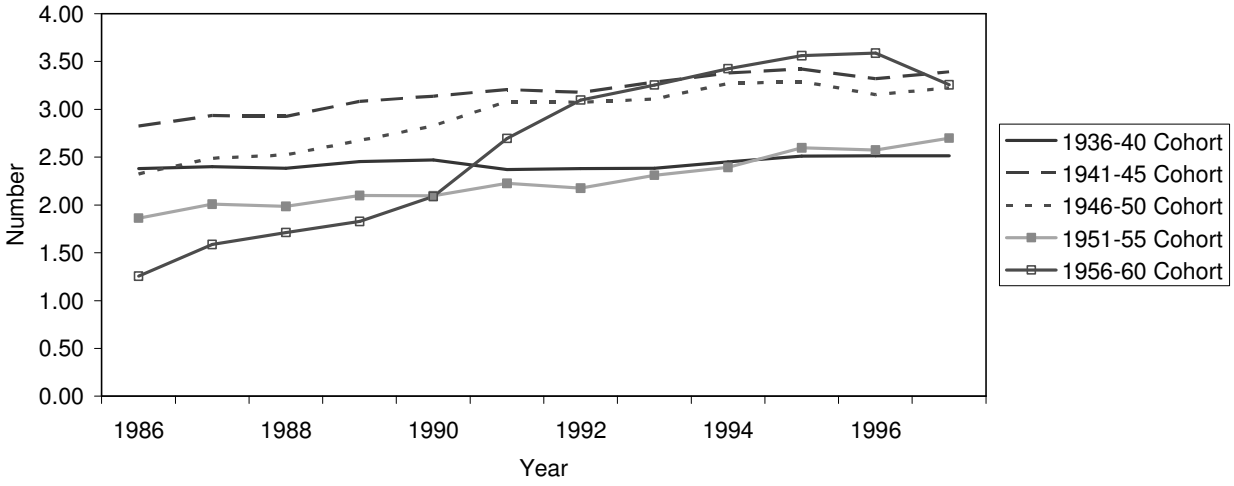


Figure 6
F-tests for Cohort, Age, and Period Models
Actual Data

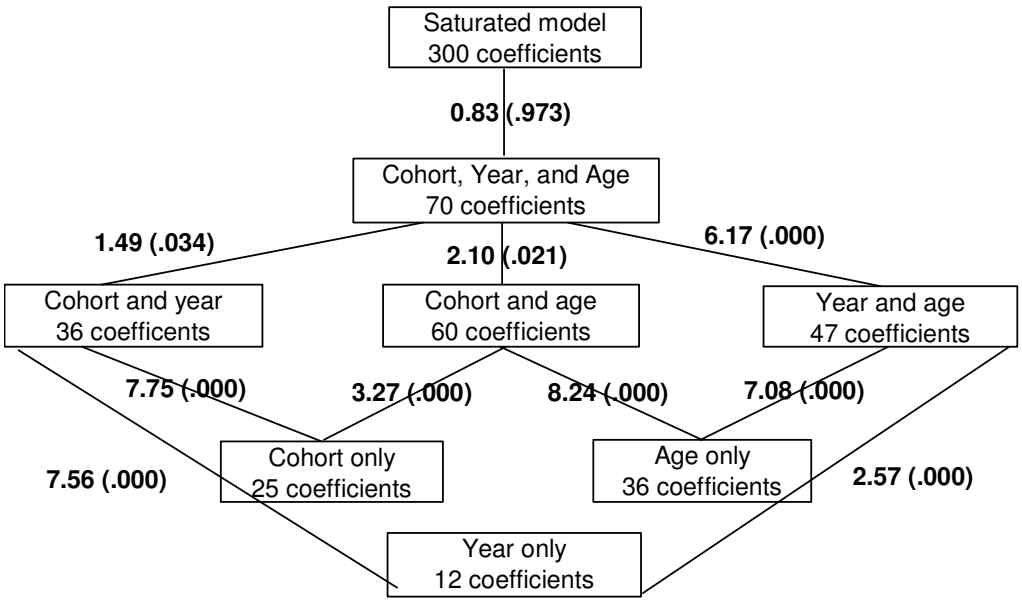


Figure 7
Geometric Average of Number of Articles Published by Age

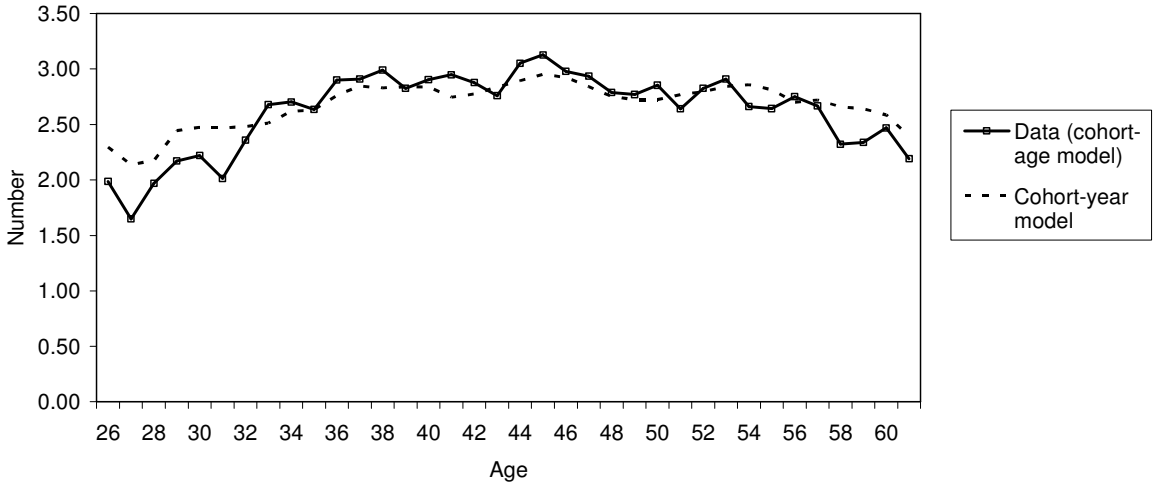


Figure 8
Geometric Average of Number of Articles Published by Year

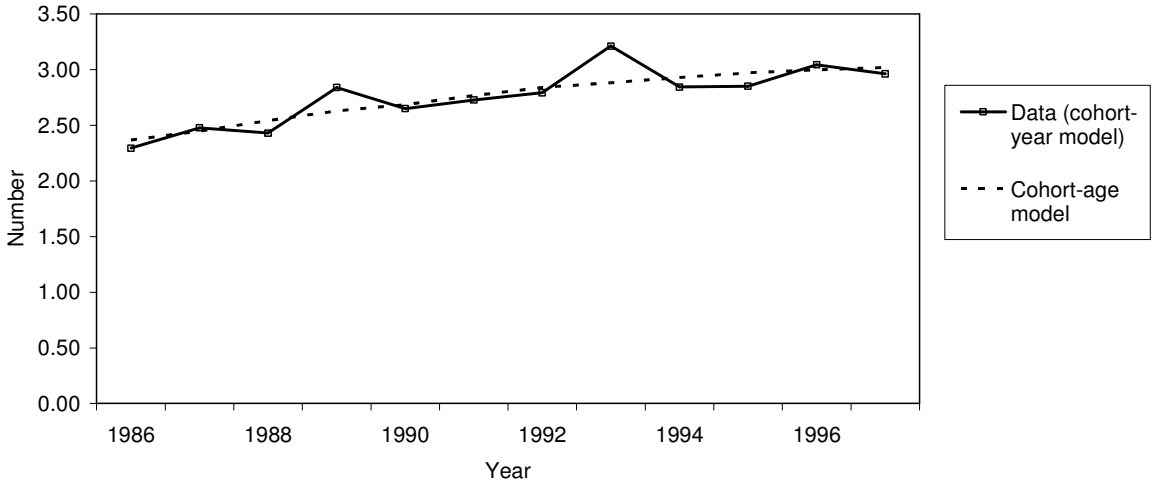
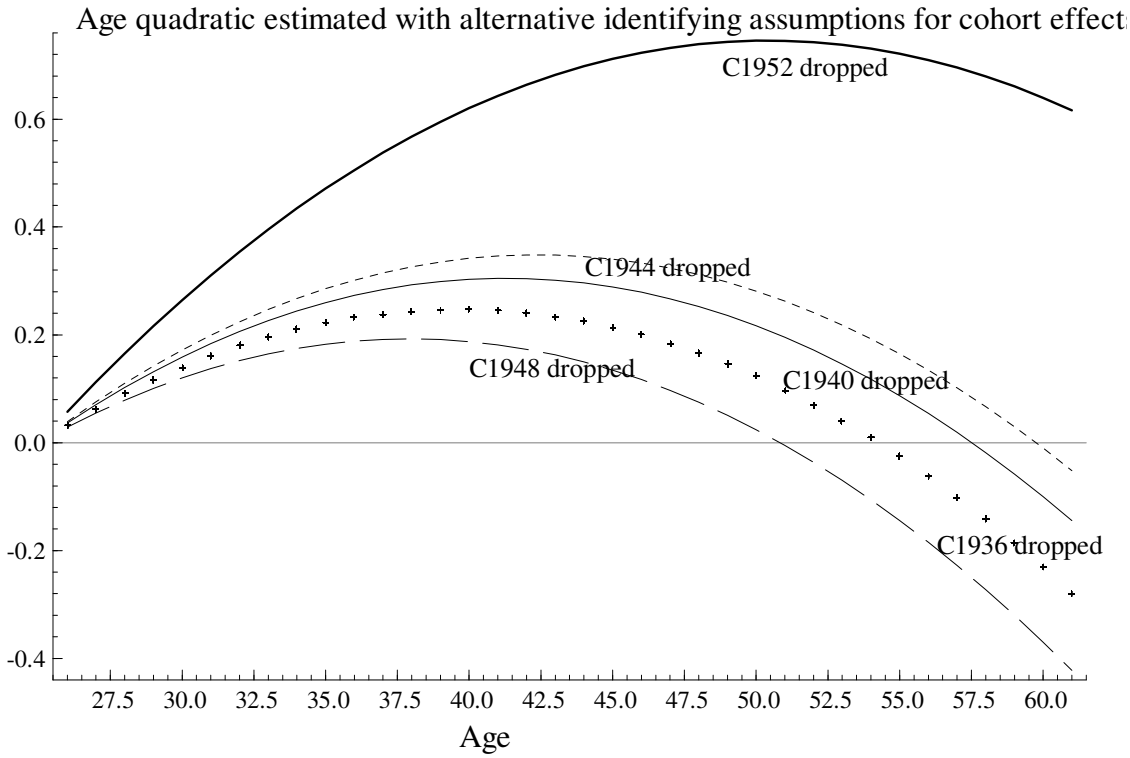


Figure 9


Age quadratic estimated with alternative identifying assumptions for cohort effect
C1952 dropped
C1944 dropped
C1948 dropped
C1940 dropped
C1936 dropped
0.6
0.4
0.2
0.0
-0.2
-0.4
27.5
30.0
32.5
35.0
37.5
40.0
42.5
45.0
47.5
50.0
52.5
55.0
57.5
60.0
Age