Holford . 1983

BIOMETRICS 39, 311–324
June 1983

The Estimation of Age, Period and Cohort Effects for Vital Rates

Theodore R. Holford

Department of Epidemiology and Public Health, School of Medicine, Yale University,
New Haven, Connecticut 06510, U.S.A.

SUMMARY

In models for vital rates which include effects due to age, period and cohort, there is aliasing due to a linear dependence among these three factors. This dependence arises both when age and period intervals are equal and when they are not. One solution to the dependence is to set an arbitrary constraint on the parameters. Estimable functions of the parameters are invariant to the particular constraint applied. For evenly spaced intervals, deviations from linearity are estimable but only a linear function of the three slopes is estimable. When age and period intervals have different widths, further aliasing occurs. It is assumed that the number of deaths in the numerator of the rate equation has a Poisson distribution. The calculations are illustrated with data on mortality from prostate cancer among nonwhites in the U.S.

1. Introduction

Time trends of incidence and mortality rates for a particular disease often provide an epidemiologist with important clues for disease etiology. Three time factors which are often considered in such an investigation are: (i) age; (ii) date of diagnosis, which we call 'period'; and (iii) date of birth, or 'cohort'. Various combinations of these three factors may be considered, and sometimes one combination provides a particularly clear summary of the data.

Frost (1939) considered the implications of these three factors on mortality rates from tuberculosis in Massachusetts. Using a graphical approach, he found that the age and cohort factors provided a consistent pattern in the trends, which was not apparent for the age and period factors. This approach was described further by Case (1956) but the technique remained a graphical one and the contributions of factors were determined visually.

While a plot provides an excellent first step in the analysis of such data, it does not provide a simple summary of the results. Model fitting can yield useful summaries of the data in terms of parameters in the model. However, a number of authors including Sacher (1960), Barrett (1973, 1978) and Fienberg and Mason (1979) have pointed out that arbitrary constraints must be applied when considering all three factors simultaneously. We shall discuss estimable functions of the parameters which are invariant as to the particular constraint applied.

Other examples of age–period–cohort analysis are found in sociology. Fienberg and Mason (1979), for example, considered the proportion of individuals finishing high school in the U.S. We shall use data on mortality from prostate cancer in the U.S. from 1935 to 1969 to illustrate the calculations.

2. Equally Spaced Age and Period Intervals

The data to be considered are a set of age-specific rates given for several periods of time. It is assumed in this section that the interval widths for age and period are equal, thus if periods

Key words: Estimable functions; Log-linear models.

are divided into five-year intervals, so is age. Age groups are represented by $i\ (=1, \ldots, I)$ and periods by $j\ (=1, \ldots, J)$. The birth cohort is defined by the age of a subject and the date of occurrence of the event of interest. Because age and period are expressed as intervals, the birth cohorts are intervals as well, but may be longer and may overlap to some extent. For example, if we have five-year age and period intervals, then individuals aged 50–54 who died during the period 1960–1964 were born sometime during the years 1905–1914. Similarly, individuals aged 55–59 belong to the 1910–1919 cohort which overlaps the previous cohort. We shall refer to a particular cohort by the index $k\ (=1, \ldots, K)$, and this is related to age and period indices by

$$k = j + I - i. \tag{2.1}$$

The rates for Age i, Period j and Cohort k are $\lambda_{ijk} = m_{ijk}/T_{ijk}$, where m_{ijk} is the expected number for the numerator of our rate, and T_{ijk} represents the person-years experience which is known. In practice, we often use the midperiod population which in most practical applications can be regarded as proportional to person-years experience. The observed number of events is represented by n_{ijk} and we shall regard these observed numbers as arising from a Poisson distribution (Armitage, 1966) with mean m_{ijk}.

In the model that we shall use, it is assumed that each factor has an additive effect on the log rate,

$$\log \lambda_{ijk} = \mu + \alpha_i + \pi_j + \gamma_k, \tag{2.2}$$

where the age effects are represented by α_i, the period effects by π_j, and the cohort effects by γ_k. We shall apply the usual constraints $\sum_i \alpha_i = \sum_j \pi_j = \sum_k \gamma_k = 0$. Sacher (1960), Barrett (1973, 1978) and Fienberg and Mason (1979) have pointed out that yet another constraint is necessary due to the linear relationship between i, j and k given in (2.1). Hence, if both age and cohort effects are in the model, the number of degrees of freedom for testing $H_0: \pi_j = 0$ for all j is $J - 2$, instead of the usual $J - 1$. Similar changes occur for tests of the age and cohort effects.

Because of the interdependency that occurs when age, period and cohort effects are considered simultaneously, a generalized inverse must be used in solving the set of normal equations that provide maximum likelihood estimators. The particular generalized inverse does not influence the significance test for parameters, but the arbitrary selection of the inverse can have a large effect on the parameters themselves. One solution is to choose a model with only two of the effects incorporated, for example, an age–period model or an age–cohort model (Baltes, 1968). Day and Charney (1981) have described a modification of this approach, which calls for simultaneous consideration of data from several sources when selecting a best subset model.

Simultaneous estimation of age, period and cohort effects has been described by Barrett (1973, 1978) and Fienberg and Mason (1979). These methods introduce an arbitrary restriction on the parameters, in that two of the effects are equated. Ideally the investigator should have a valid reason for this restriction, but it necessarily must be derived from outside the data, which may not be possible. In the remainder of this section we shall consider estimable functions of the effects, which are invariant with respect to the particular generalized inverse selected. Hence, these functions will not depend on the particular constraint used.

2.1 *Parameterization for Time Effects*

One method of characterizing the effects of an interval variable like time is to describe the trend in two components: linear trend and curvature or deviations from linearity. To illustrate, let us consider the factor 'age' represented by the effects α_i ($\sum_i \alpha_i = 0$). The linear trend can be described by the contrast

$$\alpha_L = C \sum_i c_i \alpha_i, \quad = \frac{SPD_{c,\alpha}}{SSD_c} \tag{2.3}$$

This is the regression of α_i on i ! intersection at $\bar{i}$

a p $c = p - a$

Table 1

Design matrix for equally spaced age and period intervals $I = J = 3$

i	j	k	A_C	P_C		C_C		A_L	P_L	$C_L = P_L - A_L$
1	1	3 0	1	1	−2	0	6	−1	−1	0
	2	4 1	1	−2	−1	−2	−4	−1	0	1
	3	5 2	1	1	2	1	1	−1	1	2
2	1	2 −1	−2	1	−1	2	−4	0	−1	−1
	2	3 0	−2	−2	−2	0	6	0	0	0
	3	4 1	−2	1	−1	−2	−4	0	1	1
3	1	1 −2	1	1	2	−1	1	1	−1	−2
	2	2 −1	1	−2	−1	2	−4	1	0	−1
	3	3 0	1	1	−2	0	6	1	1	0

where $c_i = i - \frac{1}{2}I - \frac{1}{2}$ and $C = (\sum_i c_i^2)^{-1}$. The curvature component is given by the age effects with the linear trend removed.

$\tilde{f}(a)$ $\cancel{a}$ $f(a) - (a - \bar{a})\cdot\delta_a = f(a) - a\delta_a - \bar{a}\delta_a$

$$\tilde{\alpha}_i = \alpha_i - c_i\alpha_L. \qquad \alpha_i - \frac{c_i}{c}\alpha_L \tag{2.4}$$

$f(a)$ $g(p)$ $h(c)$ 0 at $\bar{a}$ $\bar{p}$ $\bar{c}$

In order to parameterize these two components we partition the age columns of the design matrix into these components. For linear age we use $A_L(i) = c_i$, and for curvature we use $A_{Cl}(i)$ $(l = 1, \ldots, I - 2)$, where the A_{Cl} are orthogonal to the A_L, i.e. $\sum_i A_L(i)A_{Cl}(i) = 0$. The curvature components may be found by using second- and higher-order orthogonal polynomials or by the methods given in the Appendix. In all cases, the curvature parameters are given by

$$\tilde{\alpha}_i = \sum_l A_{Cl}(i)\alpha_{Cl},$$

where α_{Cl} represents the parameter associated with Column $A_{Cl}(\cdot)$ of the design matrix. Clearly the $\tilde{\alpha}_i$ have the linear trend removed since $\sum_i A_L(i)\tilde{\alpha}_i = 0$. In a similar manner we partition the columns for period effects by using $\mathbf{P}_L$ and $\mathbf{P}_C$ and for cohort effects by using $\mathbf{C}_L$ and $\mathbf{C}_C$; this yields the parameters π_L, π_C, γ_L and γ_C, respectively.

Using this parameterization we form the overall design matrix for the data,

$$\mathbf{X} = (\mathbf{1}\ \mathbf{A}_C\mathbf{P}_C\mathbf{C}_C\mathbf{A}_L\mathbf{P}_L\mathbf{C}_L), \tag{2.5}$$

where the row contains the regressor variables for the appropriate combination of i, j and k. Parameters corresponding to this design matrix are $\boldsymbol{\beta}' = (\mu, \boldsymbol{\alpha}'_C, \boldsymbol{\pi}'_C, \boldsymbol{\gamma}'_C, \alpha_L, \pi_L, \gamma_L)$. The matrix in (2.5) is not of full column rank because, from (2.1), we see that

$$\mathbf{C}_L = \mathbf{P}_L - \mathbf{A}_L. \tag{2.6}$$

Hence, the solution must employ generalized inverses.

As an example, consider the simple case where there are three age and period groups $(I = J = 3)$, giving $K = 5$. If we use orthogonal polynomials of degree greater than one (Fisher and Yates, 1963) to represent the curvature component then the design matrix is that given in Table 1.

2.2 *Estimable Functions of Parameters*

Using the definition given by Searle (1971, §5.4), we say that the linear function, $\mathbf{q}'\boldsymbol{\beta}$, of our parameters is estimable if $\mathbf{q}'\boldsymbol{\beta} = \mathbf{t}\log\boldsymbol{\lambda}$ for any $\mathbf{t}$, where $\log\boldsymbol{\lambda}$ is our vector of log rates. The property of particular interest here is the invariance of estimable functions to the particular parameterization used. It we use least squares, estimable functions are also best linear unbiased estimates (BLUEs) but if we use maximum likelihood estimators for Poisson

$q(X'X)^{-}(X'X) = q$

$1\times p$ $p\times p$ $1\times p$

random variables this property is not relevant. In order to test for estimability, it is sufficient to see if $\mathbf{q}'\mathbf{H} = \mathbf{q}'$, where $\mathbf{H} = \mathbf{G}\ \mathbf{X}'\mathbf{X}$, and where $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$ (Searle, 1971, p. 185).

For our design matrix (2.5) we have identified the linear dependency among the last three columns; hence, we form the partition, $\mathbf{X} = (\mathbf{X}_1 \mid \mathbf{C}_L)$. The upper left-hand portion of

$$\mathbf{X}'\mathbf{X} = \left[\begin{array}{c|c} \mathbf{X}_1'\ \mathbf{X}_1 & \mathbf{X}_1'\ \mathbf{C}_L \\ \hline \mathbf{C}_L'\ \mathbf{X}_1 & \mathbf{C}_L'\ \mathbf{C}_L \end{array}\right]$$

has a simple inverse; hence, we may use the generalized inverse

$$\mathbf{G} = \left[\begin{array}{c|c} (\mathbf{X}_1'\ \mathbf{X}_1)^{-1} & \mathbf{0} \\ \hline \mathbf{0} & 0 \end{array}\right]$$

which gives

$$\mathbf{H} = \left[\begin{array}{c|c} \mathbf{I} & (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{C}_L \\ \hline \mathbf{0} & 0 \end{array}\right].$$

The upper right-hand portion of $\mathbf{H}$ has the same form as the parameter estimates that can be used to predict $\mathbf{C}_L$ from $\mathbf{X}_1$. Using the linear dependency in (2.6), we find

$$\mathbf{H} = \left[\begin{array}{c|c} \mathbf{I} & \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \\ 1 \end{pmatrix} \\ \hline \mathbf{0} & 0 \end{array}\right]. \tag{2.7}$$

To test whether the linear trend with age is estimable we use $\mathbf{q}' = (0 \ldots 0\ 1\ 0\ 0)$ and $\mathbf{q}'\mathbf{H} = (0 \ldots 0\ 1\ 0\ 1) \neq \mathbf{q}'$. Hence, linear trend with age is not an estimable function; similarly we can show that linear trends with period and cohort are not estimable, as demonstrated by Fienberg and Mason (1979). On the other hand, $\alpha_L + \pi_L$ is estimable, and in general any function of the three linear components which has the form $d_1\alpha_L + d_2\pi_L + (d_2 - d_1)\gamma_L$ with arbitrary d_1 and d_2 is estimable. For curvature components, any linear function given by $(\mathbf{q}_c^{*\prime}, 0\ 0\ 0)\boldsymbol{\beta}$ with arbitrary $\mathbf{q}_c^*$ is estimable. Likewise the intercept is not estimable!

2.3 *Example*

To illustrate the calculations we consider data on prostate cancer mortality among nonwhites in the U.S. from 1935 to 1969; these data have also been discussed by Ernster, Selvin and Winkelstein (1978). For comparison, in the figures we show results of similar analyses of whites. The data presented in Table 2 give the number of prostate cancer deaths (National Center for Health Statistics, 1937–1973) and the estimated midperiod population for the period 1935–1960 (Grove and Hetzel, 1968) and for 1960–1969 (Bureau of the Census, 1974). A simple summary of these data is provided by the direct adjusted rates, shown in Fig. 1, which use the total male population for 1950 as the standard. Of some concern is the steady increase in deaths for nonwhites which contrasts with a modest decline for whites over the same period.

Table 2
Number of prostate cancer deaths and midperiod population for nonwhites in the U.S. by age and period

Age	Period						
	1935–	1940–	1945–	1950–	1955–	1960–	1965–
				Prostate cancer deaths			
50–	177	271	312	382	321	305	308
55–	262	350	552	620	714	649	738
60–	360	479	644	949	932	1292	1327
65–	409	544	812	1150	1668	1958	2153
70–	328	509	763	1097	1593	2039	2433
75–	222	359	584	845	1192	1638	2068
80–	108	178	285	475	742	992	1374
				Midperiod population ($\times 10^3$)			
50–	301	317	353	395	426	473	498
55–	212	248	279	301	358	411	443
60–	159	194	222	222	258	304	341
65–	132	144	169	210	230	264	297
70–	76	94	110	125	149	180	197
75–	37	47	59	71	91	108	118
80–	19	22	32	39	44	56	66

The cohort intervals for the data in Table 2 are 1850–1860, 1855–1865, 1860–1870, . . . , 1910–1920, which shall be referred to by their midpoints, 1855, 1860, 1865, . . . , 1915.

A summary of the results of fitting models to these data is given by the likelihood-ratio statistic, G^2, in Table 3. Because age is well-recognized as an important factor in cancer we do not consider models that exclude the age effect but, in principle, the need for age could also be evaluated. All of the chi square values given in Table 3 are significant at the .001 level; this is partly due to the large numbers of deaths which enable us to detect even small departures from the model. The proportion of the lack of fit that is not explained by age is high for the age–period–cohort and age–cohort models, as shown by R_A^2. In fact, when we compare the observed and fitted rates (Fig. 2) we see that even the age–cohort model provides a satisfactory fit to the data.

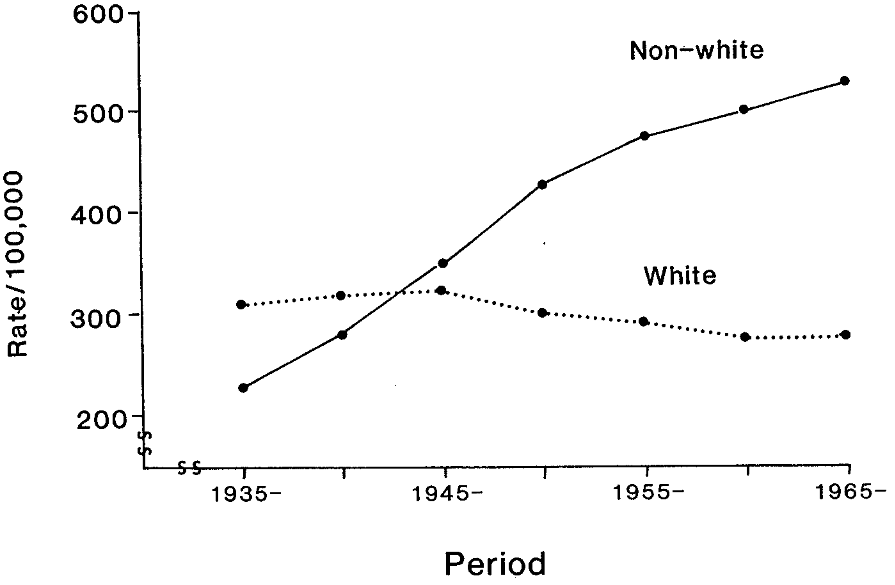

Figure 1. Age-standardized mortality rates for prostate cancer in the U.S. by race.

Table 3
Summary chi square of age (A), period (P) and cohort (C) models for data in Table 2

Model	df	G^2	R_A^2	Δdf	ΔG^2	Effect
A, P, C	25	98.91	0.97	5	28.47	$P\|A, C$
				11	622.52	$C\|A, P$
A, P	36	721.43	0.75	6	2191.92	$P\|A$
A, C	30	127.38	0.96	12	2785.97	$C\|A$
A	42	2913.35	—			

Parameters derived from fitting the age–cohort model for nonwhites are shown in Table 4. This table would provide an adequate summary of the data when a subset of the factors is used. However, we might instead consider all three factors together so that our parameters will reflect any possible adjustment for the period effects. The estimable functions are the deviations from linearity shown in Table 4 for age and cohort effects. We might also choose to include a linear component, but here we run into the estimability problem and we must make an arbitrary choice. One possibility is to assume that period is linear, $\pi_L = 0$. In this case the linear age is actually an estimate of $\alpha_L + \pi_L$, while the linear cohort is an estimate of $\gamma_L + \pi_L$. Hence, the linear age and cohort parameters are biased by whatever the true value of π_L happens to be. By adding the linear contributions shown in Table 4, we obtain effects similar to the age–cohort model. The advantage of these estimates is that they provide an adjustment for any deviations from linearity due to period.

A plot of the cohort effects is shown in Fig. 3. This suggests that in recent cohorts there is actually a decline in prostate cancer deaths for nonwhites. In fact, the main difference in the pattern of cohort effects between the races appears to be that whites exhibit a peak about

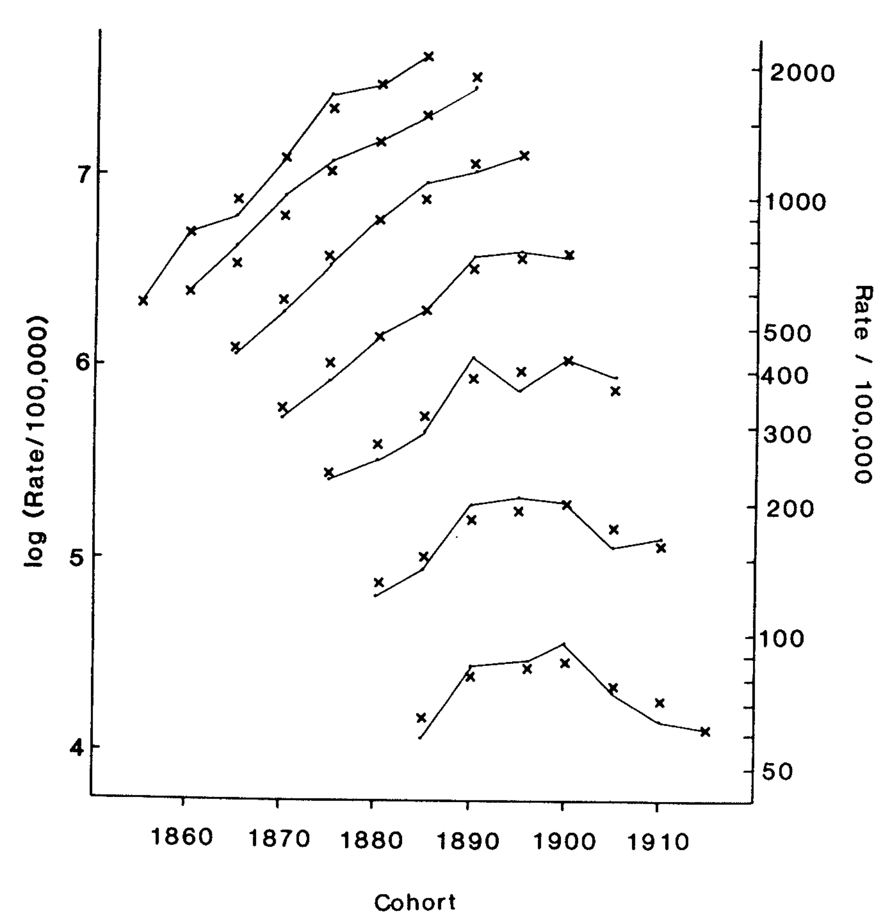

Figure 2. Observed (·) and age–cohort fitted (×) rates versus cohort by age for nonwhites.

Table 4
Age and cohort effects for age–cohort and age–period–cohort model

Group	Age–cohort	Age–period–cohort	
		Deviation	Deviation + linear
Age: 50–	−1.97	−0.242	−1.99
55–	−1.15	0.001	−1.17
60–	−0.42	0.162	−0.42
65–	0.17	0.167	0.17
70–	0.72	0.145	0.73
75–	1.17	0.013	1.18
80–	1.49	−0.245	1.51
$\hat{\alpha}_L + \hat{\pi}_L$	—	0.584	—
Cohort: 1855	−1.07	−0.376	−1.07
1860	−0.70	−0.125	−0.70
1865	−0.54	−0.092	−0.55
1870	−0.29	0.028	−0.32
1875	−0.06	0.141	−0.09
1880	0.09	0.172	0.06
1885	0.24	0.223	0.22
1890	0.44	0.320	0.44
1895	0.49	0.258	0.49
1900	0.53	0.192	0.54
1905	0.40	−0.039	0.42
1910	0.30	−0.230	0.35
1915	0.17	−0.470	0.22
$\hat{\gamma}_L + \hat{\pi}_L$	—	0.115	—

30 years earlier than nonwhites. For period we obtain deviations from linear parameters (−0.025, −0.022, 0.015, 0.054, 0.028, −0.023, −0.028) for nonwhites which are plotted in Fig. 4, together with values for whites. For nonwhites we note that while the magnitudes of the effects are small, their pattern tends to be concave downward. The cohort and period effects taken together suggest an improvement in the prostate cancer mortality rates for nonwhites;

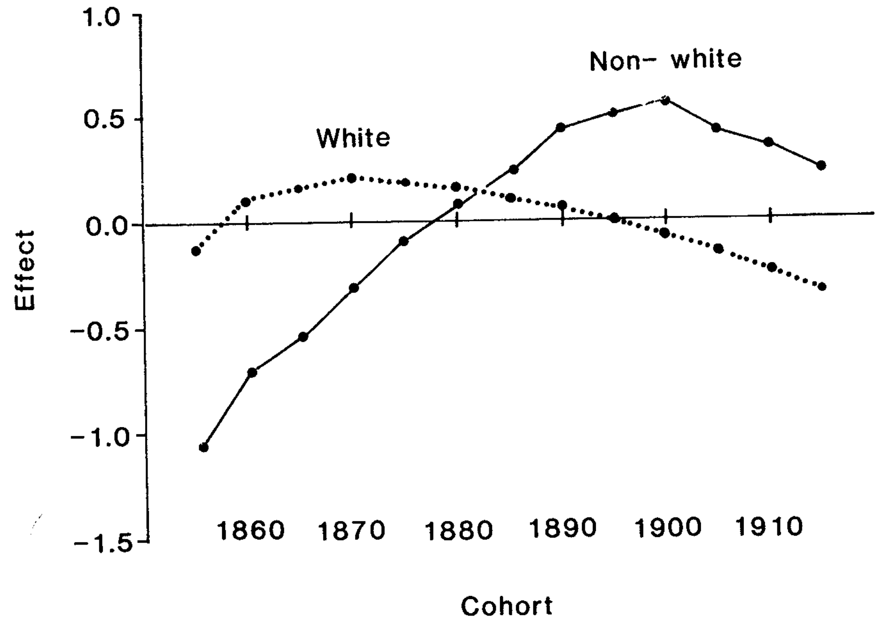

Figure 3. Cohort effects from the age–period–cohort model by race.

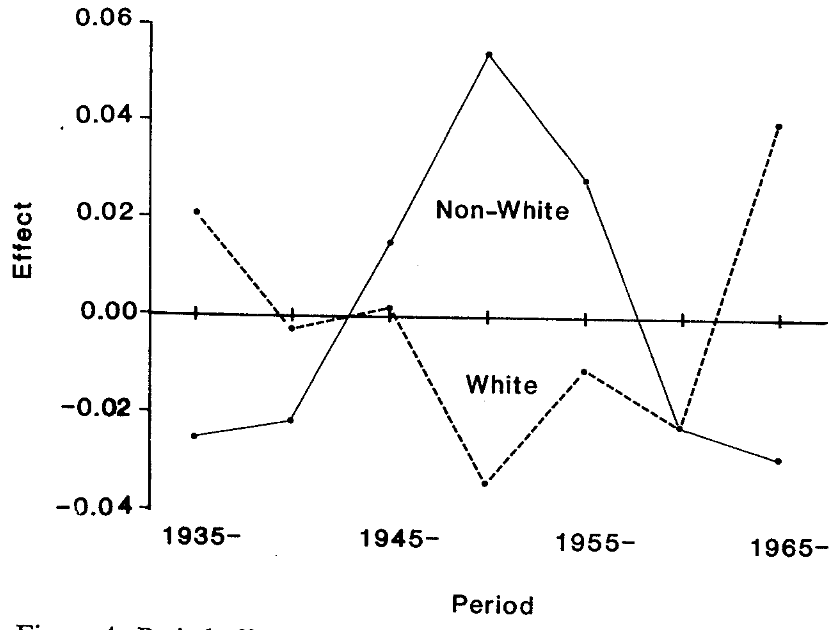

Figure 4. Period effects from the age–period–cohort model by race.

this is very different from the picture presented by the direct adjusted rates. If present trends continue, one would expect the improvement to be reflected ultimately in the age-standardized rates.

3. Balanced Unequally Spaced Intervals

Sometimes the age intervals are not equal to the period intervals. For example, rates may be reported for 10-year periods and five-year age intervals. In this section we shall consider the case in which period intervals are R times wider than age intervals. The case in which age intervals are wider than period intervals presents no new problems so it will not be discussed.

A change of one unit for the period index j $(=1, \ldots, J)$ represents a length of time equal to a change in age of R units. We shall divide age into groups for reference purposes and refer to a particular age interval by the double index (i, r), where $i = 1, \ldots, I$ and $r = 1, \ldots, R$. The design we shall consider is balanced, so the total number of age intervals is IR. This grouping of age intervals also applies to cohort intervals, which are represented by (k, r) where $k = 1, \ldots, K$ and $r = 1, \ldots, R$. As before, we have

$$k = j + I - i; \qquad (3.1)$$

and in Table 5, an example is given of these indices when $I = J = 3$ and $R = 2$. It is important to notice that under this indexing system we go from lower to higher age intervals as we increase first r then i. However, the sequence of cohorts is obtained by a decrement of r and an increment of k.

Serendipity might have resulted in the estimability problem vanishing with this change in design, but this is hardly the case. As was pointed out by Fienberg and Mason (1979), we retain the old difficulty and bring in yet another. Hence, we shall parameterize the factors affected by the grouping, in this case age and cohort, somewhat differently than in §2.

3.1 *Parameterization of Grouped Time Effects*

The grouping of the age and cohort intervals must be taken into account in the parameterization. We break down the effect of each of these factors into four components: (i) subgroup, (ii) linearity, (iii) parallelism, (iv) subgroup curvature.

Table 5
Cohort indices for unequally spaced age and period intervals where $I = J = 3$ and $R = 2$

Age	Period		
	1	2	3
(1, 1)	(3, 1)	(4, 1)	(5, 1)
(1, 2)	(3, 2)	(4, 2)	(5, 2)
(2, 1)	(2, 1)	(3, 1)	(4, 1)
(2, 2)	(2, 2)	(3, 2)	(4, 2)
(3, 1)	(1, 1)	(2, 1)	(3, 1)
(3, 2)	(1, 2)	(2, 2)	(3, 2)

For age, the subgroup variable is defined by

$$A_{Gh}(i, r) = \begin{cases} 1, & h = r, \\ -1, & h = R, \\ 0, & \text{otherwise}, \end{cases} \tag{3.2}$$

for $h = 1, \ldots, R - 1$. The linear trend among age groups is given by $A_L^*(i, r) = i - \frac{1}{2}I - \frac{1}{2}$; however, because we have an ordering on the subgroups as well, it is more natural to use

$$A_L(i, r) = R(i - \tfrac{1}{2}I - \tfrac{1}{2}) + (r - \tfrac{1}{2}R - \tfrac{1}{2}) \tag{3.3}$$

which involves the ordering among all age intervals. The quantity in (3.3) is a linear function of $A_L^*(i, r)$ and $A_{Gh}(i, r)$, so in our design matrix we will only include columns defined in (3.2) and (3.3).

Parallelism terms are given by

$$\begin{aligned} A_{Ph}(i, r) &= A_{Gh}(i, r)A_L^*(i, r) \\ &= \begin{cases} A_L^*(i, r), & h = r, \\ -A_L^*(i, r), & h = R, \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{3.4}$$

where $h = 1, \ldots, R - 1$. These terms are orthogonal to the group, as well as to the linear terms, in that $\sum_{i,r} A_L(i, r)A_{Ph}(i, r) = \sum_{i,r} A_{Gh}(i, r)A_{Ph}(i, r) = 0$.

Finally, we define the subgroup curvature component by

$$A_{Chl}(i, r) = \begin{cases} A_{Cl}(i), & h = r \\ 0, & \text{otherwise}, \end{cases} \tag{3.5}$$

where $h = 1, \ldots, R$, $i = 1, \ldots, I - 2$, and $A_{Cl}(i)$ is defined to be orthogonal to the linear component as in §2.

Hence, we have split up the $IR-1$ degrees of freedom (df) for age into contributions due to

Source	df
Group	$R - 1$
Linear	1
Parallelism	$R - 1$
Curvature	$(I - 2)R$.

Table 6 gives an example of the regressor variables for the case where $I = 3$ and $R = 2$.

The effect due to cohort is broken down in a similar manner to the age effect. The cohort group component $C_{Gh}(i, r)$ is equal to $G_{Gh}(i, r)$. Period, on the other hand, only has linear and curvature components, due to the lack of grouping. Because of the relationship among

Table 6
Regressor variables for grouped age when $I = 3$ and $R = 2$

Age interval	Group	Linear	Parallelism	Curvature	
				1	2
(1, 1)	−1	−2.5	1	1	0
(1, 2)	1	−1.5	−1	0	1
(2, 1)	−1	−0.5	0	−2	0
(2, 2)	1	0.5	0	0	−2
(3, 1)	−1	1.5	−1	1	0
(3, 2)	1	2.5	1	0	1

indices given in (3.1), the grouped linear components are related by

$$\mathbf{C}_L^* = \mathbf{P}_L - \mathbf{A}_L^* \tag{3.6a}$$

and

$$\mathbf{C}_L = R\mathbf{P}_L - \mathbf{A}_L. \tag{3.6b}$$

The entire design matrix is given by

$$\mathbf{X} = (\mathbf{1}\ \mathbf{A}_{CP}\mathbf{P}_C\mathbf{C}_{CP} \,\vdots\, \mathbf{A}_L\mathbf{P}_L\mathbf{C}_L \,\vdots\, \mathbf{A}_G \,\vdots\, \mathbf{C}_G), \tag{3.7}$$

where $\mathbf{A}_{CP}$ and $\mathbf{C}_{CP}$ are matrices which include both curvature and parallelism components. Corresponding parameters are

$$\boldsymbol{\beta}' = (\mu\ \boldsymbol{\alpha}_{CP}'\boldsymbol{\pi}_C'\boldsymbol{\gamma}_{CP}'\alpha_L\pi_L\gamma_L\boldsymbol{\alpha}_G'\boldsymbol{\gamma}_G').$$

To summarize the deviation from linearity for age, we use $\mathbf{A}_{CP}\boldsymbol{\alpha}_{CP}$ which is orthogonal to a linear age trend as well as to groups. Cohort deviations from linearity are found in a similar way.

3.2 *Estimable Functions of Parameters*

To find estimable functions of parameters we partition the design matrix $\mathbf{X}$ as shown in (3.7). Using the method outlined in §2.2, we find

$$\mathbf{H} = \left[\begin{array}{c|c|c|c} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & R \\ 0 & 0 & 0 \end{bmatrix} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{I} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array}\right].$$

Once again we see that none of the linear trends are estimable, but we can estimate $d_1R\alpha_L + d_2\pi_L + (d_2 - d_1)R\gamma_L$ for arbitrary d_1 and d_2. However, we cannot estimate the group coefficients for either age (α_{Gh}, $h = 1, \ldots, R - 1$) or cohort (γ_{Gh}, $h = 1, \ldots, R - 1$): we can only estimate the difference, $\alpha_{Gh} - \gamma_{Gh}$. The latter estimability problem arises not only for a model which includes all three effects but also for the subset model which includes only age and cohort effects.

3.3 *Example*

To illustrate the models considered for unequally spaced intervals we shall use the data in Table 2 and modify it by dropping the period 1935–1939 and the age group 80–84. For the

Table 7
Summary chi square of age (A), period (P) and cohort (C) models for unequally spaced intervals

Model	df	G^2	R_A^2	Δdf	ΔG^2	Effect
A, P, C	3	13.66	0.99	1	14.30	$P\mid A, C$
				7	409.40	$C\mid A, P$
A, P	10	423.06	0.68	2	899.39	$P\mid A$
A, C	4	27.96	0.98	8	1294.49	$C\mid A$
A	12	1322.45				

remaining data we sum the cells for adjacent periods, which leaves us with data for 10-year periods (1940–1949, 1950–1959, 1960–1969). In the modified data we group age and cohort by twos ($R = 2$) and by the number of possible levels for age groups and periods ($I = J = 3$). Hence, the relationship among indices is as shown in Table 5. This modification of the original observations effectively throws away some of the data and is only made here to show the computations involved if just the reduced data were available.

A summary of the likelihood-ratio chi square is given in Table 7; it shows a significant lack of fit for the age–period–cohort model ($p < .01$) but the R_A^2 is nearly 1 so the model does account for a great deal of the variability. Table 8 gives a summary of the deviations from linearity for all three factors. The coefficient of the subgroup variable is not included here, resulting in the sum of the odd and even deviations for both age and cohort being 0.

Estimates of linear trends are $\hat{\alpha}_L + \hat{\pi}_L = 0.6252$ and $\hat{\gamma}_L + \hat{\pi}_L = 0.0717$. If any period linear trend is ignored, the trends shown in Table 8 are obtained. A plot of the age trends is shown in Fig. 5.

A final parameter gives $\hat{\alpha}_G - \hat{\gamma}_G = 0.0010$ which is an estimable function of the subgroup parameters. The implication of this is that we may set $\hat{\alpha}_G = \Delta$ and $\hat{\gamma}_G = \Delta - 0.0010$ for an arbitrary Δ. In Fig. 5 we show the effect of this arbitrary parameter on the age effects and we

Table 8
Age, period and cohort effects for unequally spaced intervals in prostate cancer mortality among nonwhites in the U.S.

Effect		Deviation	Deviation + linear
Age:	50–	−0.160	−1.72
	55–	0.049	−0.89
	60–	0.118	−0.19
	65–	0.105	0.42
	70–	0.042	0.98
	75–	−0.154	1.41
Period:	1940–	−0.017	—
	1950–	0.033	—
	1960–	−0.017	—
Cohort:	1867.5	−0.171	−0.494
	1872.5	−0.149	−0.400
	1877.5	−0.028	−0.208
	1882.5	0.091	−0.016
	1887.5	0.159	0.123
	1892.5	0.219	0.255
	1897.5	0.189	0.296
	1902.5	0.149	0.328
	1907.5	−0.147	0.104
	1912.5	−0.310	0.013

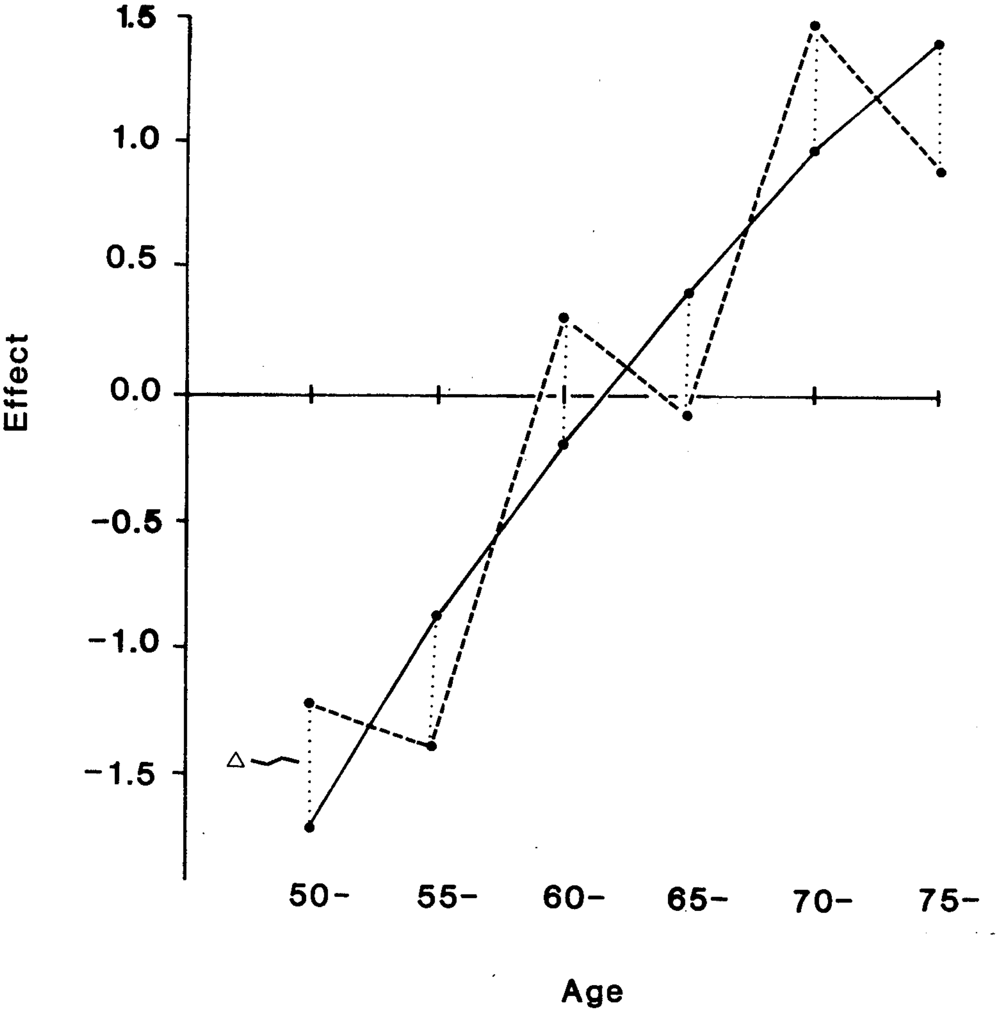

Figure 5. Age effects from age–period–cohort model for unequally spaced intervals by zero (solid line) and nonzero (broken line) Δ.

see that if it is large enough it can induce a saw-tooth shape in the trend. Such a shape seems very unnatural, but the dilemma we are in with these data is that such a phenomenon cannot be ruled out when the data are in this form. Our summary in Table 8 effectively assumes that the contribution of groups is 0, which to some extent oversmooths the trend. The fact that our estimate of $\alpha_G - \gamma_G$ is small indicates that whatever group effect does exist, there would need to be a corresponding cancelling out between age and cohort, which is implausible.

4. Discussion

A model which assumes that, on some scale, there is an additive effect due to age, period and cohort is in itself arbitrary. We might instead have considered interactions, but in fact if we look at interactions among any two factors, the third factor spans a subspace of that interaction space. One view of this model is that it is between a two-factor main effect model and a model that includes all interaction terms which would be saturated. In some instances a simpler summary of the data might be obtained by using a two-factor model with interactions.

The models fitted here require a generalized inverse for their solution and so an arbitrary constraint must be used. Barrett (1973, 1978) and Fienberg and Mason (1979) suggested equating certain parameters; this would introduce an arbitrary element which can have a large effect on the parameter estimates and their trend. L. L. Kupper and J. M. Janis, in an unpublished report (Institute of Statistics, University of North Carolina, Mimeo Series

No. 1311, 1980), used principal components to avoid the linear dependence in the design matrix and, in effect, to introduce another set of arbitrary constraints. In this paper we have identified a set of estimable functions which are invariant to the actual constraints used. These functions are not unique, but they have been chosen to give one summary of the results which can be easily interpreted. This approach also identifies which component of the parameters cannot be estimated.

In these examples we have used maximum likelihood estimators for the parameters, assuming a Poisson error distribution. The significant lack of fit is due primarily to the large number of deaths rather than to systematic departures from the model. This is not very surprising when dealing with rates in large populations. In such instances, fitting the log rates by means of least squares would not be unreasonable. On the other hand, when frequencies are not so large, the significance level of the goodness-of-fit statistic would be critical and maximum likelihood would be preferable to least squares.

The conclusions reached by this analysis of the prostate cancer data are substantively different from those suggested by the direct adjusted rates. This is another instance where summary rates can smooth out important features in the data when the appropriate model does not hold (Freeman and Holford, 1980). Model fitting can lead to an understanding of the factors that are important, and a better summary of the data can often be obtained by using the fitted model.

Acknowledgements

Some of this work was carried out while the author was on leave at the Department of Biomathematics, University of Oxford, supported by an Eleanor Roosevelt International Fellowship from the International Union Against Cancer. The technical assistance of Nancy Amery is greatly appreciated.

Résumé

Lorsque les modèles portant sur des statistiques vitales incluent des effects dus à l'âge, à la période et à la cohorte, il y a confusion venant de la dépendance linéaire entre ces trois facteurs. La dépendance intervient que les intervalles sur l'âge et la période soient égaux ou non. Une solution à l'étude de cette dépendance est d'imposer une contrainte sur les paramètres. Les fonctions estimables des paramètres ne dépendent pas de la contrainte particulière utilisée. Pour des intervalles de temps, on peut estimer la non linéarité seulement grâce à une fonction linéaire des trois pentes. Quand les intervalles d'âge et de période sont différents, d'autres confusions interviennent. On suppose que le nombre de morts au numérateur de l'équation donnant le taux est poissonnien. Les calculs sont illustrés à l'aide de données de mortalité due au cancer de la prostate chez les non-blancs des Etats-Unis.

References

Armitage, P. (1966). The chi-square test for heterogeneity of proportions after adjustment for stratification. *Journal of the Royal Statistical Society, Series B* **28,** 150–163.

Baltes, P. B. (1968). Longitudinal and cross-sectional sequences in the study of age and generation effects. *Human Development* **11,** 145–171.

Barrett, J. C. (1973). Age, time and cohort factors in mortality from cancer of the cervix. *Journal of Hygiene* **71,** 253–259.

Barrett, J. C. (1978). The redundant factor method and bladder cancer mortality. *Journal of Epidemiology and Community Health* **32,** 314–316.

Bureau of the Census (1974). Current population reports. *Population Estimates and Projections.* Series P-25, No. 519. Washington, D.C.: U.S. Government Printing Office.

Case, R. A. M. (1956). Cohort analysis of mortality rates as an historical or narrative technique. *British Journal of Preventive and Social Medicine* **10,** 159–171.

Day, N. E. and Charnay, B. (1982). Time trends, cohort effects and ageing as influence on cancer incidence. In *Trends in Cancer Incidence,* K. Magnus (ed.), 51–65. Washington: Hemisphere.

Draper, N. R. and Smith, H. (1966). *Applied Regression Analysis.* New York: Wiley.

Ernster, V. L., Selvin, S. and Winkelstein, W. (1978). Cohort mortality for prostatic cancer among United States nonwhites. *Science* **200,** 1165–1166.

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

Fisher, R. A. and Yates, F. (1967). *Statistical Tables for Biological, Agricultural and Medical Research.* London: Oliver and Boyd.

Freeman, D. H. and Holford, T. R. (1980). Summary rates. *Biometrics* **36,** 195–205.

Frost, W. H. (1939). The age selection of mortality from tuberculosis in successive decades. *American Journal of Hygiene, Section A* **30,** 91–96.

Grove, R. D. and Hetzel, A. M. (1968). *Vital Statistics Rates in the United States* 1940–1960. Washington, D.C.: National Center for Health Statistics.

National Center for Health Statistics (1937–1973). *Vital Statistics of the United States* (1935–1969). Washington, D.C.: U.S. Government Printing Office.

Sacher, G. A. (1960). Analysis of life tables with secular terms. *American Institute of Biological Sciences Symposium* **6,** 253–257.

Searle, S. R. (1971). *Linear Models.* New York: Wiley.

Received January 1982

Appendix

In order to generate the curvature component of evenly spaced interval variables, we generate a set of regressor variables which are orthogonal to the linear trend. One method is to use orthogonal polynomials but many tables do not go above a fifth degree so higher-degree terms must be generated by a formula. An alternative set of regressors can be found by making the usual analysis-of-varianc design matrix orthogonal to the linear term. If we have I levels of an equally spaced ordinal variabl then the usual regressor variables can be given in a matrix **Z**, where the ith row and jth column given by

$$Z_{ij} = \begin{cases} 1, i = j, \\ -1, i = I, \\ 0, \text{otherwise} \end{cases}$$

with $i = 1, \ldots, I$ and $j = 1, \ldots, I - 1$.

Linear trend can be found by using the column vector **L**, where

$$L_i = i - \tfrac{1}{2}I - \tfrac{1}{2}.$$

Using the method described by Draper and Smith (1966, p. 156), we can generate a matrix orthogonal to **L**,

$$\mathbf{Z}^* = \mathbf{Z} - \mathbf{L}(\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}\mathbf{Z}.$$

The elements of this matrix are given by

$$\mathbf{Z}^*_{ij} = \begin{cases} M_{ij} + 1, & i = j, \\ M_{ij} - 1, & i = I, \\ M_{ij}, & \text{otherwise,} \end{cases}$$

where

$$M_{ij} = \{-12\ L_i(L_j - \tfrac{1}{2}I + \tfrac{1}{2})\}/\{I(I - 1)(I + 1)\}.$$

To avoid a linear dependency we only use the first $I - 2$ columns of $\mathbf{Z}^*$, denoted by $\mathbf{Z}^*_0$. The curvature component would be found from $\mathbf{Z}^*_0\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ are the parameters associated with the columns of $\mathbf{Z}^*_0$.