ESTIMATION OF TIME, AGE, AND COHORT EFFECTS

R. M. Portman, H. L. Lucas, R. C. Elston, and B. G. Greenberg

This investigation was supported in part by Public Health Service Research Grant No. RG-9605 from the Division of General Medical Sciences and by the Public Health Service Training Grant

Institute of Statistics
Mimeographed Series No. 372
April, 1963
<table>
<thead>
<tr>
<th>Page</th>
<th>Line</th>
<th>Errata</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>8</td>
<td>advice instead of advise (misspelled)</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>States Public Health (omission)</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>obtain minimum-variance (omission of hyphen)</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
<td>the $z_{ij}$ of the introduction (subscript)</td>
</tr>
<tr>
<td>19</td>
<td>9</td>
<td>$f_1(w) f_2(w) \ldots f_g(w)$ (underline vector)</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>as for (3.3) (not (3.2))</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>such as (3.2) (not (3.2))</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>$\theta_o$ (subscript)</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>$\theta_n$ (subscript)</td>
</tr>
<tr>
<td>21</td>
<td>14</td>
<td>$\ldots$ of $C^*$ form a basis (misspelled)</td>
</tr>
<tr>
<td>22</td>
<td>9</td>
<td>Scheffe' (accent)</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Scheffe' (accent)</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>$\ldots$ u-factor model $\mathbb{E}(y)$ (omission)</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>$V(G^<em>) \subseteq V^+(C^</em>)$, (omission of line under $C$)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>$W(r \times r + 1) = (-1</td>
</tr>
<tr>
<td>25</td>
<td>13</td>
<td>$G^*$ (omission of *)</td>
</tr>
<tr>
<td>26</td>
<td>7</td>
<td>$\ldots$ then $R(A)$ would (misspelled)</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>$G^*$ (omission of *)</td>
</tr>
<tr>
<td>31</td>
<td>18</td>
<td>suppose that the (misspelled)</td>
</tr>
<tr>
<td>38</td>
<td>14</td>
<td>$\ldots$ relation of cohort (omission)</td>
</tr>
<tr>
<td>53</td>
<td>16</td>
<td>$\Sigma \Sigma z_{ij}^r / n$ (subscripts)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\Sigma \Sigma w_{ij}^s / n, \Sigma \Sigma z_{ij}^r / n$ (subscripts)</td>
</tr>
<tr>
<td>63</td>
<td>1</td>
<td>segments with (plural)</td>
</tr>
<tr>
<td>68</td>
<td>8</td>
<td>$\alpha_u^*$ (omission of star)</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>$\beta_v^*$ (omission of star)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>$\gamma_g^*$ (omission of star)</td>
</tr>
<tr>
<td>74</td>
<td>26</td>
<td>Scheffe' (accent)</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

| LIST OF TABLES                          | v          |
| LIST OF FIGURES                        | vi         |
| CHAPTER I - INTRODUCTION               | 1          |
| CHAPTER II - REVIEW OF LITERATURE      | 8          |
| CHAPTER III - CRITERIA FOR ESTIMABILITY| 15         |
| Preliminary Considerations             | 15         |
| Notation and Definitions               | 15         |
| The Model in Matrix Form               | 16         |
| The Problem                            | 18         |
| The Modified Model                     | 19         |
| The u-Factor Model                     | 19         |
| The Contrasts of Interest              | 21         |
| Useful Lemmas                          | 22         |
| Conditions for Estimability with the u-Factor Model | 22         |
| Remarks                                | 24         |
| Value of the Theorem                   | 24         |
| The Cohort Model                       | 25         |
| CHAPTER IV - PRACTICAL PROBLEMS OF ESTIMATION | 27         |
| Definition of Basic Cohorts            | 27         |
| Effect on the Estimates of Modifications of the Model | 30         |
| CHAPTER V - EXAMPLES                   | 35         |
| Data Used                              | 35         |
| Methods of Analysis                    | 36         |
| Results of Study I                     | 40         |
| Results of Study II                    | 54         |
| Discussion and Conclusions             | 65         |
| CHAPTER VI - SUMMARY AND RECOMMENDATIONS | 68         |
| Resume of Results                      | 68         |
| Recommendations                        | 71         |
| Suggestions for Future Work           | 72         |
| LIST OF REFERENCES                     | 74         |
LIST OF TABLES

5.1. Age specific discovery rates of primary and secondary syphilis among negro females per 1,000 population, 1941 - 1947, by age and year of discovery in study area ........ 35

5.2. Age-specific "incidence" rates for breast cancer in females per 100,000 population, Connecticut, 1935 - 1951 ........ 36

5.3. Table to show which basic cohorts are combined in each particular modification of the model - study I ........ 41

5.4. Estimates of the mean and mean square error - study I .... 55

5.5. Table to show which basic cohorts are combined in each particular modification of the model - study II .... 56

5.6. Estimates of the mean and mean square error - study II ... 64
LIST OF FIGURES

5.1. Estimated relative age effects for study I, modifications
      1-1 through 1-4 .................................. 42
5.2. Estimated relative time effects for study I, modifications
      1-1 through 1-4 .................................. 43
5.3. Estimated relative cohort effects for study I, modifications
      1-1 through 1-4 .................................. 43
5.4. Estimated relative age effects for study I, modifications
      2-1 and 2-2 ...................................... 44
5.5. Estimated relative time effects for study I, modifications
      2-1 and 2-2 ...................................... 45
5.6. Estimated relative cohort effects for study I, modifications
      2-1 and 2-2 ...................................... 46
5.7. Estimated relative age effects for study I, modifications
      3-1 and 3-2 ...................................... 47
5.8. Estimated relative time effects for study I, modifications
      3-1 and 3-2 ...................................... 48
5.9. Estimated regression on cohorts for study I, modifications
      3-1, 3-2, and 4-1 ................................. 49
5.10. Estimated regression on ages for study I, modifications
      4-1 and 4-2 ...................................... 50
5.11. Estimated relative time effects for study I, modifications
      4-1 and 4-2 ...................................... 51
5.12. Estimated relative age effects for study II, modifications
      1-1 through 1-9 .................................. 57
5.13. Estimated relative time effects for study II, modifications
      1-1 through 1-9 .................................. 57
5.14. Estimated relative cohort effects for study II, modifications
      1-1 through 1-9 .................................. 58
5.15. Estimated relative age effects for study II, modifications
      2-1 and 2-2 ...................................... 59
5.16. Estimated relative time effects for study II, modifications
      2-1 and 2-2 ...................................... 59
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.17</td>
<td>Estimated relative cohort effects for study II, modifications 2-1 and 2-2</td>
<td>60</td>
</tr>
<tr>
<td>5.18</td>
<td>Estimated relative age effects for study II, modifications 3-1, 3-2, and 3-3</td>
<td>61</td>
</tr>
<tr>
<td>5.19</td>
<td>Estimated relative time effects for study II, modifications 3-1, 3-2, and 3-3</td>
<td>61</td>
</tr>
<tr>
<td>5.20</td>
<td>Estimated regression on cohorts for study II, modifications 3-1, 3-2, and 3-3</td>
<td>62</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

The control of disease is of concern to many individuals and agencies, firstly for immediate humanitarian reasons and secondly to combat the long-range effects of disease on the national economy. Efficient control requires knowledge of the factors affecting the incidence and the prevalence of a disease and the rate of mortality resulting therefrom. The study of factors relating to disease presents a complex problem owing to the many possible causal elements. With the growth of biological and social sciences over the last century, an increasing number of such factors are being recognized. The only source of knowledge of these factors is past experience, predominantly in the form of vital statistics. At the present time, it is not possible to isolate the effect of each possible factor because of a lack of information about, and practical restrictions on, the collection of data. There is, however, almost always the possibility of studying the effects of factors in three groups, viz. those related to time, those associated with age, and finally those related to the generation, or cohort. The effects of factors related to a cohort are confounded with the interaction of the first two. It is, therefore, important to study methods of retrieving information about the effects of these three groups of factors.

The rates which constitute the data fall into a natural rectangular array ordered by age intervals and time intervals. A third natural classification occurs in this array, namely the generation or cohort.
The cohort is most readily identified when the intervals of age and time are all equal. If the age intervals and the time intervals are represented in the array of data by rows and columns, respectively, the cohorts, which consist of individuals born in the same time interval, are represented by diagonals. Members of a given cohort would be expected to experience a more homogeneous environment throughout life than members of different cohorts. In the problem considered here, the intervals for age and time are assumed to be consecutive, but not necessarily of equal length. When they are all of equal length, the classification is referred to as the symmetric case.

Many biological phenomena have been found empirically to be represented well by a relationship of the type: \( z = ae^{cx} \). Furthermore, age, time, and cohorts are believed to affect the rates in a multiplicative way. On this premise we shall develop a model which appears to be reasonable for characterizing the data.

Assume that a given rate applies to a rectangular cell which has been obtained by combining adjacent rows and columns from a more refined classification which is a symmetric case; e.g.

Designate this cell as the \( ij^{th} \) cell and a given cell of the refined classification of the \( uv^{th} \) sub-cell.
If the refined classification is fine enough, we can regard each sub-cell as containing only one age effect (the $u^{th}$), only one time effect (the $v^{th}$), and only one cohort effect (the $g^{th}$). Thus the model for the $uv^{th}$ sub-cell is

$$y_{uvw}^* = \mu + \alpha_u^* + \beta_v^* + \gamma_g^* + \epsilon_{uvw}$$

where $y_{uvw}^* = \log (r_{uvw})$ and $r_{uvw}$ is the rate of the $uv^{th}$ sub-cell.

To obtain the rate for the $ij^{th}$ cell, compute

$$y_{ij}^* = \frac{1}{n_{ij}} \sum_u \sum_v \sum_g y_{uvw}^*$$

where the limits of summation are proper for the $ij^{th}$ cell, and $n_{ij}$ is the number of sub-cells in the $ij^{th}$ cell.

Now let

- $n_{iju}$ = the number of sub-cells in the $i_j^{th}$ cell that contains the $u^{th}$ effect
- $n_{ijv}$ = the number of sub-cells in the $i_j^{th}$ cell that contains the $v^{th}$ effect.
- $n_{ijg}$ = the number of sub-cells in the $i_j^{th}$ cell that contains the $g^{th}$ cohort effect.

Note that

$$\sum_u n_{iju} = \sum_v n_{ijv} = \sum_g n_{ijg} = n_{ij}.$$  \hspace{1cm} (1.1)

Then

$$y_{ij}^* = \mu + \frac{\sum_u n_{iju}}{n_{ij}} \alpha_u^* + \frac{\sum_v n_{ijv}}{n_{ij}} \beta_v^* + \frac{\sum_g n_{ijg}}{n_{ij}} \gamma_g^* + \epsilon_{ij}.$$  \hspace{1cm} (1.2)
where
\[
\varepsilon_{ij} = \frac{\sum \sum \sum \varepsilon_{uvg}}{n_{ij}}.
\]

We may write (1.2) as
\[(1.3) \quad y_{ij}^* = u + \sum u w_{uij} \alpha_u^* + \sum v x_{vij} \beta_v^* + \sum g z_{gij} \gamma_g^* + \varepsilon_{ij}\]

and it is apparent from (1.1) and (1.2) that
\[(1.4) \quad \sum_u w_{uij} = \sum_v x_{vij} = \sum_g z_{gij} = 1.\]

In general, \(i = a, a - 1, \ldots, 1,\) with \(a\) being the youngest age interval, and \(j = 1, 2, \ldots, t,\) with \(t\) being the most recent time interval.

It is now apparent that cells which are not rectangular, but which are polygons in shape, will have shapes that can be resolved into several rectangles adjoined; e.g.

\[
\text{The rate for such a cell can be computed as an appropriate weighted mean of the several rectangles involved. Thus (1.3) and (1.4) apply regardless of the shape of the cell.}
\]

Note that in special cases, the model can be simplified. In the symmetric case for example, we can define
\[(1.5) \quad \alpha_i = \sum_u w_{uij} \alpha_u^*\]
because the $w_{uij}$ are the same for all $j$. Similarly, we define

\[(1.6) \quad \beta_j = \sum_v x_{vij} \beta_v^* \]

because the $x_{vij}$ are the same for all $i$, and finally we define

\[(1.7) \quad \gamma_k = \sum_g z_{gij} \gamma_g^* \]

because the $z_{gij}$ are the same for all $i,j$ satisfying $k = i + j - 1$.

Another simple case is the non-symmetric one in which the lengths of the age intervals are not equal and/or the lengths of the time intervals are not equal. In this case (1.5) and (1.6) still hold, but (1.7) does not.

Finally, it may be desirable to replace the step-function type parameters used in the foregoing development for age and/or cohorts by the parameters of a smooth function, e.g., the parameters of a polynomial on $u$ and/or $g$. For the linear term, the independent variables for the $ij^{th}$ cell are:

\[\hat{w}_{ij} = \sum_u (u) w_{uij} \quad \text{for age} \]

and

\[\hat{z}_{ij} = \sum_g (g) z_{gij} \quad \text{for cohort.} \]

For the quadratic term, the coefficients are:

\[\hat{w}_{ij}^2 = \sum_u (u^2) w_{uij} \quad \text{for age} \]

and

\[\hat{z}_{ij}^2 = \sum_g (g^2) z_{gij} \quad \text{for cohort,} \]

etc.
In the development given

\[ y_{ij}^* = \frac{1}{n_{ij}} \sum_{u} \sum_{v} \sum_{g} \log r_{uvg} . \]

In practice, however, the \( r_{uvg} \) are never known. Instead, the rate data are of the form,

\[ r_{ij} = \frac{1}{n_{ij}} \sum_{u} \sum_{v} \sum_{g} r_{uvg} \]

and one would use

\[ y_{ij} = \log r_{ij} \]

as an approximation. If the \( r_{uvg} \) are fairly uniform over the \( ij^{th} \) cell, the discrepancy between \( y_{ij}^* \) and \( y_{ij} \) will not be very great. Furthermore, this discrepancy becomes less important if it is necessary to use \( y_{ij} = \log (r_{ij} + h) \), where \( h \) is some positive constant, in order that \( y_{ij} > -\infty \).

For rates specified insofar as possible as to sex, race, geographic location, etc., the functional relationship of (1.3) may give both a meaningful and an adequate model for a specific disease, at least over a limited age range. This model would not be especially useful in predicting trends because extrapolation outside the time region covered by the study would be involved.

The value of using this approach in the study of rates is conceived to be in revealing causal factors. Suppose it is possible to obtain estimates of all linear contrasts of the type:
\( \alpha_i - \alpha_{i'} \), \( i \neq i' \)

\( \beta_j - \beta_{j'} \), \( j \neq j' \)

\( \gamma_k - \gamma_{k'} \), \( k \neq k' \)

in model (1.3). Then the basic age pattern of the disease can be evaluated, and time and cohort effects can be separately studied in conjunction with social and medical history for clues to more specific factors influencing the rates.

Since model (1.3) is linear in the parameters, general linear estimation theory may be used to obtain minimum variance, unbiased estimates of all estimable linear functions of the parameters. The problem considered here is the determination of the conditions this model must satisfy for all contrasts of interest to be estimable. "All contrasts of interest" will be taken to mean those contrasts indicated in (1.8) or linear combinations of them. Consideration will also be given to the properties of the estimates and to the problems inherent in dealing with actual data. When a smooth-function representation of the effects of a factor is employed, the actual parameters of the function, rather than contrasts among them, will be designated as contrasts of interest.
CHAPTER II

REVIEW OF LITERATURE

The study of mortality rates has a long history. Early attempts to describe mortality experience for the purposes of prediction were based on the assumption that the mortality curve is a "smooth" function of age. This assumption appears to be reasonable since the cumulative total mortality experience of a human population approximates a strictly monotonic function of age.

Jordan [7] gives Abraham de Moivre credit for the earliest proposal (1724) of a mathematical expression relating survival to age. De Moivre assumed the relationship to be linear and claimed to obtain a good approximation for the age range of 12 to 86 years.

Benjamin Gompertz [5] proposed in 1825 that the force of mortality increases essentially in geometric progression for the age range of about 10 to 60 years. Although his function, \( L_x = d g^x \), where \( L_x \) = number living at age \( x \), is based only on age, he suggested that death may be the result of two factors, one being deterioration and the other chance. This relationship is still employed by actuaries.

Characterizations of the mortality curve have also been made directly from observed age-specific, standardized death rates in the form of life tables. It has been noted that these rates vary with sex, race, state of urbanization, and nation. Such life tables also shift with time. Moreover, most tables in use are not longitudinal; i.e., each rate is based on a different population representing an age group born in a different era.
In 1934, Kermack, et al., [8] published an extensive study of mortality rates in Scotland, England, and Sweden. The results supported their hypothesis that the predominant factor determining the health of an individual for life is environment up to the age of about 15 years. They suggested that the age-specific death rate is proportional to a function of age multiplied by a function of the year of birth.

Greenwood [4] discussed the prediction of mortality using age, time, and generation, and concluded that it was unreasonable to predict future rates on the basis of only one of these factors. He investigated methods of graduation and prediction from a combination of two factors, age and time, or age and generation, to fit Makeham's modification of Gompertz's Law of Life,

\[ \mu_x = A + B \ e^x. \]

Using a minimum \( \chi^2 \) method for estimating the parameters, A and B, from years or generations, he concluded that the use of either combination did not greatly distort the facts but that neither method produced good predictions as judged by a \( \chi^2 \) goodness-of-fit test.

In 1939, Frost [2] introduced the term "cohort" to indicate the generation factor. His study of sex-age-specific tuberculosis death rates in Massachusetts revealed a remarkable parallelism of cohort contours, but the time contours showed no consistency and could lead to misinterpretation.

Case [1] recommended cohort contour analysis of disease-sex-age-specific mortality and morbidity rates in conjunction with social and medical history as a general narrative technique to make inferences from the
rates. He introduced the idea of considering the number of years which individual cohorts contribute to each datum or rate. For example, consider a rate figure, \( r_{ij} \), for the \( j^{\text{th}} \) time period of three years and the \( i^{\text{th}} \) age interval of three years. There are five one-year cohorts associated with this rate. The central one contributes three years of experience to the rate, the two adjacent cohorts contribute two years each, and the two extreme cohorts contribute one year each. This procedure generalizes for the case in which the age and time intervals of the rates cover \( n \) units of time, \( n \) being an odd integer. In this instance, \( 2n - 1 \) cohorts of one time unit each contribute to the rate. Of these, the central cohort contributes \( n \) units of experience to the rate, while the cohorts on either side contribute successively fewer, and the two extremes contribute only one unit of time each. Case further suggested that mortality be considered to be a product of three factors: one which is biological and inescapable (age); one which is attributable to early nurture (cohort); and one which is the result of constantly changing environment and therapy (time).

In analyzing the incidence of syphilis, Greenberg, et al., [3] considered age, time, and cohort as the primary factors affecting the race-age-sex-specific rates in the four county area studied. The model used was:

\[
\begin{align*}
\gamma_{xij} + 1 &= x_i \cdot e^{\beta_2 + \beta_0 x_0 + \beta_1 x + \sum_{i=1}^{7} \gamma_i x_i + \sum_{j=8}^{14} \delta_j x_j + \varepsilon_{xij}}
\end{align*}
\]  

(2.1)

where
\[ x_0 = 1 \text{ for all observations} \]
\[ x = \text{age} \ (15 \leq x \leq 29) \]
\[ x_1 = 1 \text{ in 1941, 0 otherwise} \]
\[ \ldots \]
\[ x_7 = 1 \text{ in 1947, 0 otherwise} \]
\[ x_8 = 1 \text{ for those born 1912-1914, 0 otherwise} \]
\[ \ldots \]
\[ \ldots \]
\[ x_{14} = 1 \text{ for those born 1930-1932, 0 otherwise} \]

and \[ \beta_0, \beta_1, \beta_2, \] are parameters in a Pearson Type III curve,
\[ e = \text{base of natural logarithms} \]
\[ \gamma_i = \text{year effects} \]
\[ \delta_j = \text{date of birth or cohort effects} \]
\[ \epsilon_{xij} = \text{random error, assumed normally and independently distributed with mean 0 and variance } \sigma^2 \]

and \[ y_{xij} = \text{rate for age } x \text{ in year } i \].

They used a least-squares procedure to estimate the meaningful parameters after making the logarithmic transformation. They noted that in this analysis the cohorts are a part of the interaction between time and age. The assumption of normally distributed, independent errors was examined and found to be reasonable, and the small estimate of residual variance indicated a good fit of the model. Age, year, and cohort effects were found to be significant in the statistical sense.
Methods of interpretation were indicated as well as instances for which erroneous inferences would have been possible had the observed rates been analyzed for each factor separately.

Hussein [6] used a similar model to analyze the incidence of breast cancer in Connecticut women. It was:

\[
\log e y_{ijk} = \mu x_0 + \sum_{i=1}^{9} \alpha_i x_i + \sum_{j=10}^{21} \beta_j x_j + \sum_{k=22}^{25} \gamma_k x_k + \epsilon_{ijk}
\]

where

\[
x_0 = 1 \text{ for all observations}
\]

\[
x_1 = \begin{cases} 
1 & \text{for observations in age 30-34} \\
0 & \text{otherwise}
\end{cases}
\]

\[
x_9 = \begin{cases} 
1 & \text{for observations in age group 70-74} \\
0 & \text{otherwise}
\end{cases}
\]

\[
x_{22} = \begin{cases} 
1 & \text{if observation lies in year 1935-1937} \\
0 & \text{otherwise}
\end{cases}
\]

\[
x_{25} = \begin{cases} 
1 & \text{if observation lies in year 1948-1951} \\
0 & \text{otherwise}
\end{cases}
\]

\[
x_{10} = \begin{cases} 
2_1(ij) & \text{if observation is affected by cohort 1} \\
0 & \text{otherwise}
\end{cases}
\]
\( x_{21} = \begin{cases} z_{12}(ij) & \text{if observation is affected by cohort 12} \\ 0 & \text{otherwise} \end{cases} \)

\( \mu \) = true mean or intercept

\( \alpha_i \) = the age effect

\( \beta_j \) = the cohort effect

\( \gamma_k \) = the year effect

\( \epsilon_{ijk} \) = the random error term, which is assumed to be normally and independently distributed with mean 0 and a common variance \( \sigma^2 \)

and \( \nu_{ijk} \) = the incidence rate observed in the \( i^{th} \) age group, the \( j^{th} \) cohort, and the \( k^{th} \) year.

The four time periods for which the data were collected varied in length from three to five years. There were 61 one-year cohorts contributing to the study. From these, he formed 11 five-year intervals and 1 six-year interval. He then assigned a value of 0, 1, ..., or 6 to the cohort coefficient \( z_k \) according to whether 0, 1, ..., or 6 one-year sub-cohorts contributed to the \( ij^{th} \) rate. Although this was not precisely the weighting scheme suggested by Case, Hussein found little difference in the sums of squares for regression for the two methods, and there was some computational advantage in his method. Note that the coefficients of the cohort effects in Hussein's model do not sum to the same quantity for all observations. From (1.3) of the introduction, it is evident that they must if the model is to be reasonable. Examination of his model showed that not all of the contrasts of interest were estimable. However, all three factors were shown to be statistically
significant under the assumption of independent, normally distributed errors, and the goodness of fit of the model was indicated by the fact that the estimated residual variance was small. For purposes of interpretation, he pointed out that the effect of a cohort or a time period is really a function of several variables for which individual values are not available, and that the grouping of the variables into one factor is especially advantageous when they are so interrelated.
CHAPTER III
CRITERIA FOR ESTIMABILITY

Preliminary Considerations

Notation and Definitions. The problem of estimability is most easily studied with vector and matrix algebra. The notation used is the following:

\( M (r \times s) \) denotes a matrix with \( r \) rows and \( s \) columns.

\( M' (s \times r) \) denotes the transpose of \( M \).

\( I \) denotes an identity matrix.

\( 0 \) denotes a null matrix.

\( \mathbf{m} (s \times 1) \) denotes a column vector of \( s \) elements and \( \mathbf{m}' (1 \times s) \) a row vector of \( s \) elements.

\( \mathbf{1} \) denotes a column vector with all elements unity.

\( \mathbf{0} \) denotes a null column vector.

\( V_s \) denotes the vector space containing all vectors of \( s \) elements.

\( V (M) \) denotes the vector space spanned by the row vectors of \( M (r \times s) \). \( V (M) \) is a subspace of \( V_s \).

\( V^+ (M) \) denotes the orthogonal complement (orthocomplement) of \( V (M) \); i.e., the vector space consisting of all vectors in \( V_s \) orthogonal to \( V (M) \). \( V^+ (M) \) may contain the null vector only. If the rows of \( P \) form a basis of \( V^+ (M) \), then \( V (P) = V^+ (M) \) and \( M P^' = 0 \).

\( R (M) \) denotes the rank of \( V (M) \).

\( R^+ (M) \) denotes the rank of \( V^+ (M) \).
\[ V(M) \supseteq V(Q) \] denotes that \( V(M) \) contains \( V(Q) \), implying that \( R(M) \geq R(Q) \).

\[ V(M) \supset V(Q) \] denotes that \( V(M) \) strictly contains \( V(Q) \), implying that \( R(M) > R(Q) \).

\[ V(M) \perp V(Q) \] denotes that \( V(M) \) is orthogonal to \( V(Q) \); the necessary and sufficient condition is that \( M Q' = 0 \).

\[ \hat{\lambda} \] denotes an estimate of \( \lambda \).

\[ E(\ ) \] denotes the expected value of any variable placed in the parentheses.

**The Model in Matrix Form.** In matrix notation, the step-function model (1.3) for age intervals, \( t \) time intervals, where (1.5) and (1.6) hold, and \( m \) cohorts can be written as:

\[
E(\gamma) = A \tau = A_0 \tau_0 + A_1 \tau_1 + A_2 \tau_2 + A_3 \tau_3
\]

where

\[ \gamma \ (n \times 1) \] is the vector of transformed rates arranged by time (1 through \( t \)) within age (1 through \( a \)); \( n = ta \)

\[ \tau \ (p \times 1) \] is the vector of parameters; \( p = 1 + a + t + m \).

\( A \ (n \times p) \) is a known matrix of coefficients.

To be meaningful, \( a, t, \) and \( m \) must each be greater than one. In (3.1) \( \tau \) is partitioned as follows:

\[
\tau' = [\tau_0' \mid \tau_1' \mid \tau_2' \mid \tau_3']
\]

where
\( \mathbf{1}_0 \ (1 \times 1) = \) the over-all mean

\( \mathbf{1}_1 \ (a \times 1) = \) the vector of age interval effects

\( \mathbf{1}_2 \ (t \times 1) = \) the vector of time interval effects

\( \mathbf{1}_3 \ (m \times 1) = \) the vector of cohort effects.

The matrix \( A \) is correspondingly partitioned: i.e.

\[
A = [A_0 \mid A_1 \mid A_2 \mid A_3]
\]

where

\[
A_0 \ (n \times 1) = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}; \text{ each } 1 \text{ is } t \times 1
\]

\[
A_1 \ (n \times a) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}; \text{ each } 1 \text{ and } 0 \text{ is } t \times 1
\]

\[
A_2 \ (n \times t) = \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix}; \text{ each } I \text{ is } t \times t
\]

\[
A_3 \ (n \times m) = \text{ the part of the } A \text{ matrix pertaining to the cohorts; its elements are the } z_g \text{ of the introduction and zeros appropriately arranged.}
\]

Note that \( A_0 \ 1 \ (1 \times 1) = A_1 \ 1 \ (a \times 1) = A_2 \ 1 \ (t \times 1) = A_3 \ 1 \ (m \times 1) = 1 \ (n \times 1) \), as was explained in the introduction.
The Problem. The problem is to find the conditions required of A for all contrasts of interest to be estimable. One complete set of contrasts of interest can be written as:

\[(3.2) \quad \mathbf{A}(p-4 \times 1) = \mathbf{C}(p-4 \times p) \mathbf{r}(p \times 1)\]

where

\[\mathbf{r}'\] is defined as for (3.1)

\[
\mathbf{C} = \begin{bmatrix}
0 & \mathbf{C}_1 & 0 & 0 \\
0 & 0 & \mathbf{C}_2 & 0 \\
0 & 0 & 0 & \mathbf{C}_3
\end{bmatrix}
\]

and \(\mathbf{C}_1\) (a \(-1 \times a\)), \(\mathbf{C}_2\) (t \(-1 \times t\)) and \(\mathbf{C}_3\) (m \(-1 \times m\)) are each of the form, \([I \mid -1]\).

A contrast among the parameters, \(\lambda' = [\lambda_1, \ldots, \lambda_w]\) is by definition a linear function of the \(\lambda_i\), \(\mathbf{r}'\lambda = \sum_{i=1}^{w} f_i \lambda_i\), with known constant coefficients subject to the condition \(\sum_{i=1}^{w} f_i = 0\), i.e., \(\mathbf{r}'\mathbf{l} = 0\). Any row, \(\mathbf{c}_i'\), of \(\mathbf{C}\) thus meets the necessary condition to be a coefficient vector for a contrast among the parameters.

The row vectors of \(\mathbf{C}\) form a basis of the set of coefficient vectors of all possible contrasts of interest. This is shown by permuting columns 1, 1 + a, and 1 + a + t of \(\mathbf{C}\) to the last three column positions. This leaves an identity matrix of order \(p-4\) in the first \(p-4\) columns, proving that the \(p-4\) rows of \(\mathbf{C}\) are linearly independent. Furthermore, the coefficient vector of any contrast of interest can be obtained by a linear combination of the rows of \(\mathbf{C}\).
The Modified Model. It may be of interest to describe age and/or cohorts by a smooth function such as a polynomial since these factors may be considered to have underlying distributions. To illustrate the modification of the model where age is to be described by a polynomial of degree $a'$, with $0 \leq a' < a$, the model is

$$E(y) = \bar{A}_T = A_0 \bar{\tau}_0 + \bar{A}_1 \bar{\tau}_1 + A_2 \bar{\tau}_2 + A_3 \bar{\tau}_3$$

where

$$\bar{\tau}_1 (a' \times 1) = \text{the vector of regression coefficients for age}$$

$$\hat{A}_1 (n \times a') = [f_1(w) f_2(w) \ldots f_{a'}(w)];$$

$f_q(w)$ is the $q$th transformation of the $w$th age and the other matrices and vectors are as defined for (3.1).

In this case, the contrasts of interest are

$$\tilde{A} (p' - 1 \times 1) = \tilde{C} \tilde{p} (p' - 1 \times p') \tilde{r} (p' \times 1).$$

Where $\tilde{r'}$ is defined as for (3.2)

$$\tilde{C} = \begin{bmatrix} 0 & I_1 & 0 & 0 \\ 0 & 0 & C_2 & 0 \\ 0 & 0 & 0 & C_3 \end{bmatrix}$$

Where $I_1$ is $a' \times a'$ and $C_2$ ($p' \times 1 \times 1$) and $C_3$ ($m - 1 \times m$) are each of the form $[I | -1]$, $p' = 1 + a' + t + m$. The problem for this model is to find the conditions of estimability for the regression coefficients of the smooth function used to describe the age factor as well as for the contrasts of interest among the step-function parameters.

The u-Factor Model. A general model will be considered, of which the cohort model (3.1) or a modification of it such as (3.2), are
special cases. A model which aside from the general mean is affected by u factors which do not interact will be called simply the "u-factor" model. In matrix notation, this model is:

\[(3.5) \quad E(\mathbf{y}) = \mathbf{B} \mathbf{\theta} = B_0 \theta_0 + B_1 \theta_1 + \cdots + B_{r-1} \theta_{r-1} + B_{r+1} \theta_{r+1} + \cdots + B_u \theta_u \]

where

\[\mathbf{y} (w \times 1) = \text{the appropriately ordered vector of observations}\]

\[\mathbf{\theta} (v \times 1) = \text{the vector of parameters}\]

\[\mathbf{B} (w \times v) = \text{the known matrix of coefficients}.\]

In \((3.5)\) \(\mathbf{\theta}\) is partitioned as follows:

\[\mathbf{\theta}' = [\theta_0' \mid \theta_1' \mid \cdots \mid \theta_{r-1}' \mid \theta_{r+1}' \mid \cdots \mid \theta_u']\]

where

\[\theta (1 \times 1) = \text{the over-all mean}\]

\[\theta (v_h \times 1) = \text{the vector of step-function effects of the } h^{th} \text{ factor; } h = 1, 2, \ldots, r\]

\[\theta_k (v_k \times 1) = \text{the vector of parameters of the smooth function for the } k^{th} \text{ factor; } k = r + 1, r + 2, \ldots, u.\]

Note that \(v = \sum_{h=1}^{r} v_h + \sum_{k=r+1}^{u} v_k + 1;\) and that \(0 \leq r \leq u.\) The coefficient matrix \(\mathbf{B}\) is correspondingly partitioned; i.e.,

\[\mathbf{B} = [B_0 \mid B_1 \mid \cdots \mid B_{r-1} \mid B_{r+1} \mid \cdots \mid B_u]\]

where

\[B_0 \text{ is } w \times 1\]

\[B_h \text{ is } w \times v_h\]

\[B_k \text{ is } w \times v_k\]
with $B_0 = B_{-h} (v_k \times 1)$ for $h = 1, 2, ..., r \leq u$, but $B_0 \neq B_{e_k}$ for $k = r + 1, r + 2, ..., u$ with $e_k = \frac{1}{1} (v_k \times 1)$ times any non-zero scalar. This will happen, for example, if the $k^{th}$ factor is quantitative in some sense (such as is age) and its effects are to be described by a polynomial of degree $v_k$, with $0 < v_k < v_k^*$ where $v_k^*$ is the number of "levels" for the $k^{th}$ factor.

The Contrasts of Interest. The basic set of contrasts of interest for the $u$-factor model will be defined as

$$\hat{a}^* (v - r - 1 \times 1) = c^* (v - r - 1 \times v) \theta (v \times 1)$$

where

$$\theta$$ is defined as for (3.5),

$$c^* = \begin{pmatrix} 0 & c_1^* & 0 & ... & 0 & 0 & ... & 0 \\ 0 & 0 & c_2^* & ... & 0 & 0 & ... & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & ... & c_r^* & 0 & ... & 0 \\ 0 & 0 & 0 & ... & 0 & I_{r+1} & ... & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & ... & 0 & 0 & ... & I_u \end{pmatrix}$$

where the $c^*_h (v_h - I \times v_h)$ are of the form $[I \mid -1]$ and the $I_k$ are $(v_k \times v_k)$. That the row vectors of $c^*$ for a basis for the set of coefficient vectors for all contrasts of interest follows by induction from the argument for $C$ of (3.2). Clearly, $R(c^*) = v - r - 1$. 
Useful Lemmas. Two lemmas necessary for the subsequent development will now be cited. The first may be found in some form in a treatise on linear estimation, and the second can be derived easily from a theorem presented in such a treatise.

Lemma 1. If $M$ is $r 	imes s$ and $R(M) = q$, $q \leq s$, then $R^+(M) = s - q$.

Lemma 2. Given that $E(y) = M\theta$, the linear functions $F\lambda$ of the parameters are estimable if and only if $V(M) \supseteq V(F)$, which implies that $V^+(F) \supseteq V^+(M)$.

Lemma 1 is a rewording of lemma 10, page 384 of Scheffe [11]. Lemma 2 is easily proved from theorem 1, page 13 of Scheffe [11]. In the context of $E(y) = X'\theta$, he states: "The parametric function $\psi = c'\theta$ is estimable if and only if $c'$ is a linear combination of the rows of $X'$, i.e. if and only if there exists a vector $a$ such that $c' = a'X'".

Conditions for Estimability with the u-Factor Model.

Theorem: Given the u-factor $E(y) = B\theta$, where

$$B(w \times v) = \begin{bmatrix} B_0 & B_1 & B_2 & \ldots & B_r & B_{r+1} & \ldots & B_u \end{bmatrix}$$

$$\theta'(1 \times v) = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 & \ldots & \theta_r & \theta_{r+1} & \ldots & \theta_u \end{bmatrix}$$

and the subscripts, 0, 1, ..., $r$, $r + 1$, ..., $u$ indicate that the numbers of columns in the matrices or the number of elements in the vectors so designated are 1, $v_1$, ..., $v_r$, $v_{r+1}$, ..., $v_u$, respectively, with $v = \sum_{h=1}^{r} v_h + \sum_{k=r+1}^{u} v_k + 1$, $0 \leq r \leq u$, and $B_0 = B_{h-1}(v_h \times 1)$ for $h = 1, 2, \ldots, r$ but $B_0 \neq B_{k-1}$ for $k = r + 1, r + 2, \ldots, u$.
where \( a_k = 1 (v_k \times 1) \) times any non-zero scalar, and given the basic set of contrasts of interest, \( \mathbf{z}^* = C^* \mathbf{g} \), where

\[
C^* (v - r - 1 \times v) = \begin{pmatrix} 0 & C^*_1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
0 & 0 & C^*_2 & \ldots & 0 & 0 & \ldots & 0 \\
0 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & \ldots & C^*_r & 0 & \ldots & 0 \\
0 & 0 & 0 & \ldots & 0 & I_{r+1} & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 0 & 0 & 0 & I_u \\
\end{pmatrix}
\]

where the \( C^*_h (v_h - 1 \times v_h) \) are of the form \([I | -I]\) and the \( I_k \) are \((v_k \times v_k)\), then \( \mathbf{z}^* \) is estimable if and only if \( V^+(B) = V (G^*) \), where

\[
G^* (r \times v) = \begin{pmatrix} -1 & 1^t & 0^t_2 & \ldots & 0^t_r & 0^t_{r+1} & \ldots & 0^t_u \\
-1 & 0^t_1 & 1^t & \ldots & 0^t_r & 0^t_{r+1} & \ldots & 0^t_u \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
-1 & 0^t_1 & 0^t_2 & \ldots & 1^t & 0^t_{r+1} & \ldots & 0^t_u \\
\end{pmatrix}
\]

Proof of theorem: From lemma 2, \( \mathbf{z}^* \) is estimable if and only if \( V^+(B) \subseteq V^+(C^*) \). Thus to prove the theorem, it is only necessary to show that \( V (G^*) \subseteq V^+(C^*) \), and that \( V^+(B) = V (G^*) \).
To show that $V(G^\ast) = V^+ (C^\ast)$, first it is shown that $V^+ (C^\ast) = V (P)$, where

$$P(r + 1 \times v) = \begin{bmatrix}
1 & 0_1 & 0_2 & \ldots & 0_r & 0_{r+1} & \ldots & 0_u \\
0 & 1_1 & 0_2 & \ldots & 0_r & 0_{r+1} & \ldots & 0_u \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0_1 & 0_2 & \ldots & 1_r & 0_{r+1} & \ldots & 0_u
\end{bmatrix}.$$  

Note that $R(C^\ast) = v - r - 1$, and from lemma 1, that $R^+ (C^\ast) = r + 1$.

Next note that $R(P) = r + 1$ and that $C^\ast P^t = 0$. Thus $V^+ (C^\ast) = V (P)$. Now $G^\ast = WP$, where $W (r \times r + 1) = [-1 \mid 1]$. Therefore, $V (G^\ast) \subseteq V^+ (C^\ast)$.

To show that $V^+ (B) = V (G^\ast)$, we first note that $V^+ (B) \subseteq V^+ (C^\ast)$, because the form of $B$ is such that $BP^t \neq 0$ and $V^+ (B) \neq V^+ (C^\ast)$. This implies that $R^+ (B) \leq r < R^+ (C^\ast) = r + 1$. Thus, by lemma 1, $R(B) \geq v - r$.

We now note that the form of $B$ is such that $B G^\ast = 0$, and, since $R(C^\ast) = r$, this implies by lemma 1 that $R(B) \leq v - r$. But, from the above, $R(B) \geq v - r$. Therefore, for $G^\ast$ to be estimable, $V^+ (B) = V (G^\ast)$, $R(B) = v - r$ and $R^+ (B) = r$.

Remarks

Value of the Theorem. The importance of this theorem is that the estimability of all contrasts of interest can now be established with greater ease than by use of lemma 2. Any linear model, whether step-function type (except with interaction), smooth-function type or a combination of the two, can be written as a special case of the u-factor
model. Furthermore, the useful contrasts to be estimated in such a model are of the type $\mathbf{c}^*$ of (3.6). Thus to insure that all of the desired estimates are obtainable, it is necessary and sufficient to establish that $R(B) = v - r$.

In some cases, it may be obvious that $R(B) < v - r$, and therefore that it is impossible to obtain all of the desired estimates without additional assumptions. In case there is any doubt that $R(B) = v - r$, however, the sweep-out method described by Rao [10] and others is a convenient method of finding $R(B)$.

The Cohort Model. In particular the cohort model (3.1), namely,

$$E(Y) = A = A_0 T_0 + A_1 T_1 + A_2 T_2 + A_3 T_3$$

is a special case of the $u$-factor model with $r = u = 3$. Because $A_0 = 1 = A_{1-h}(h = 1, 2, 3)$ the theorem applies directly and the $G$ matrix is

$$G(3 \times p) = \begin{bmatrix}
-1 & 1' & 0' & 0' \\
-1 & 0' & 1' & 0' \\
-1 & 0' & 0' & 1'
\end{bmatrix}$$

with the vectors designated by the subscripts 1, 2 and 3 containing $a$, $t$ and $m$ elements, respectively. Note that $v$ of the theorem is $p = a + t + m + 1$. Because the contrasts of interest, $\mathbf{c}$, are estimable if and only if $V^+(A) = V(G)$, estimability requires that $R(A) = p - 3$.

If $n$, the number of observations, is less than $ta$ because observations are missing in some cells, it is still necessary and sufficient
for estimability of all contrast of interest in the context of Model (3.1) to show rank of \( A = p - 3 \). The rank of a coefficient matrix which would be \( p - 3 \) were all observations available, may still be \( p - 3 \) depending on the number and location of the missing observations. Naturally, if a time effect or an age effect became completely confounded with a cohort effect by the location of the missing observations, the \( R(A) \) would be less than \( p - 3 \). Also, \( n \geq p - 3 \) must hold if \( R(A) = p - 3 \). For either case, estimability of all contrasts of interest can be attained in a variety of ways; for example, by reducing \( m \), the number of cohorts, or \( a \) or \( t \) by combining age intervals or time intervals, or some combination of these.

When the cohort model is modified so that age and/or cohort effects are represented by a smooth-function, the theorem applies with \( 1 \leq r < 3 \). For example, the \( G \) matrix for model (3.3) is

\[
(3.8) \quad \tilde{G} (2 \times p') = \begin{pmatrix} -1 & 0'_{12} & 0'_{13} \\ -1 & 0'_{21} & 0'_{23} \end{pmatrix}
\]

By reasoning analogous to that for the unmodified model, \( R(A) = p' - 2 \) and corresponding statements can be made.
CHAPTER IV

PRACTICAL PROBLEMS OF ESTIMATION

Definition of Basic Cohorts

Having established the condition on the coefficient matrix for all contrasts of interest to be estimable, it is desirable to re-examine the cohort model. In the symmetric case, the basic (step-function) model is now defined to be:

\[(4.1) \quad \mathbf{E(Y)} = A^* = A_0^* + A_1^* + A_2^* + A_3^*\]

where

\[A_0, \, A_1 \, \text{are as defined in } (3.1), \, \text{and in particular}\]

\[A_3 (n \times m) = \begin{pmatrix}
 I (t \times t) & 0 (t \times a - 1) \\
 0 (t \times 1) & I (t \times t) & 0 (t \times a - 2) \\
 \vdots & & \ddots \\
 0 (t \times a - 2) & I (t \times t) & 0 (t \times 1) \\
 0 (t \times a - 1) & I (t \times t)
\end{pmatrix}
\]

Note that \(k = i + j - 1\) and \(m = a + t - 1\).

This seems to be a natural definition for the cohorts because of the way in which rates are calculated; i.e.

\[r_{ij} = I_{ij} / M_{ij}\]

or

\[r_{ij} = I_{ij} / (M_{ij} - \frac{1}{2} I_{ij})\]
where

\[ I_{ij} \] = the number of cases (deaths) of individuals age \( i \) occurring in time interval \( j \)

\[ M_{ij} \] = the number of individuals at risk at age \( i \) in time interval \( j \), estimated at mid-interval.

Assuming that no migration is involved, the individuals contributing to \( M_{ij} \) are the survivors of the group of individuals contributing to \( M_{i-1,j-1} \) and thus are the natural cohorts, with cohort interval equal to that of time and age.

For model \((4.1)\), \( R(A) < p - 3 \). This is shown by the fact that \( A G'_1 = 0 \) and \( R(G'_1) = 4 \), where

\[ G'_1 (p \times 4) = \begin{bmatrix} G \\ G'^t \end{bmatrix} \]

where

\( G \) is defined in \((3.7)\)

\[ G'^t = [0 (x - 1)_1^t (x - 1)_2^t (x - 1)_3^t] \]

and \((x - 1)_x^t\) = the row vector whose \( x^{th} \) element is \( x - 1 \), and the subscripts 1, 2, and 3 denote that the vector contains \( a, t, \) and \( m \) elements respectively. It is easily seen that \( A G' = 0 \). To show that \( A G^* = 0 \), note that if \( e' \) is the row of \( A \) corresponding to the element \( y_{ij} \) of \( Y \), then

\[ e' G^* = (i - 1) + (j - 1) - (k - 1) \]

\[ = 1 + j - 2 - (i + j - 2) = 0. \]
That the rows of $G_1$ are linearly independent is shown by the fact that columns $2$, $a + 2$, $a + t + 2$, and $a + t + 3$ of $G_1$ form an identity matrix of rank $4$. Thus $R(G_1) = 4$.

Because $R(A) < p - 3$, not all contrasts of interest are estimable. Therefore, in order to obtain useful estimates, it is necessary to make further assumptions about the model. For example, it can be shown that, if we assume that at least two of the adjacent cohort effects are equal, so that $m \leq a + t - 2$, then $R(A) = p - 3$.

In the non-symmetric case, the problem is more complicated. There no longer exists the natural cohort population defined in the way the data are presented. This is because there are always some individuals common to the $(i - 1)^{th}$ age interval and $i^{th}$ age interval, and/or common to the $(j - 1)^{th}$ time period and the $j^{th}$ time period.

According to the development of the model in the introduction, we can form cohorts by combining sub-cohorts. Then the cohort effect on any rate is a weighted mean of those of the cohorts contributing to it. It was not possible to investigate non-symmetric cases in general. It was found in some cases, however, that the maximum number of such cohorts that can be defined and have $R(A) = p - 3$ is $a + t - 1$. In other cases fewer cohorts must be defined in order that $R(A) = p - 3$.

Although this method of forming cohorts and dealing with them makes it possible to obtain useful estimates, we note that there may be additional correlations in the errors which we cannot remove.
Effect on the Estimates of Modifications of the Model

Consider now that the following model is the appropriate one, i.e., that it represents the true state of nature:

$$\bar{Y} = A\vec{r} + \vec{\epsilon}$$

where

- $A$ = the coefficient matrix
- $\vec{r}$ = the vector of parameters
- $\vec{\epsilon}$ = the vector of random variables such that $E(\vec{\epsilon}) = \vec{0}$,
  $E(\vec{\epsilon} \vec{\epsilon}') = I$
- $\bar{Y}$ = the vector of observations.

Then, according to well known linear estimation theory, unique unbiased estimates of all linearly estimable functions of the parameters can be obtained which are best in the sense that the estimates have minimum variance of all unbiased estimates.

To indicate the effect on the estimates of an inappropriate modification of the model, we shall compare in detail two alternative models for the symmetric case. We shall start by considering model (4.1) modified so that $\gamma_{a+t-2} = \gamma_{a+t-1}$; thus the one parameter $\gamma_{a+t-2}$ will suffice for both quantities and $m = a + t - 2$. It is convenient at this point to reparameterize the model in order to obtain the contrasts given in (5.2) directly. The modified model is then:

$$E(y_{ij}) = \mu^* + \alpha_i^* + \beta_j^* + \gamma_k^*$$

where
\[ \mu^* = \mu + \alpha_a + \beta_t + \gamma_m \]

\[ \alpha_i^* = \alpha_i - \alpha_a; \quad i = 1, 2, \ldots, a - 1 \]

\[ \beta_j^* = \beta_j - \beta_t; \quad j = 1, 2, \ldots, t - 1 \]

\[ \gamma_k^* = \gamma_k - \gamma_m; \quad k = 1, 2, \ldots, a + t - 3; \quad m = a + t - 2; \]

or, in matrix notation,

\[(4.3) \quad \mathbf{E}(\mathbf{y}) = A^* \mathbf{1}^* .\]

The estimates \( \hat{\tau}^* \) of the contrasts \( \tau^* \) are obtained in the usual way by solution of the normal equations, i.e.

\[(4.4) \quad \hat{\tau}^* = (A^* A^*)^{-1} A^* \mathbf{y} .\]

To show that the estimates are unbiased estimates of the contrast of interest, if the modified model (4.3) does in fact hold, we note:

\[(4.5) \quad E(\hat{\tau}^*) = (A^* A^*)^{-1} A^* E(\mathbf{y}) = (A^* A^*)^{-1} A^* A^* \hat{\tau}^* = \hat{\tau}^* .\]

Under the same model (4.3) the variances and covariances of the estimates are given by:

\[ E(\hat{\tau}^* - \tau^*)(\hat{\tau}^* - \tau^*)' = (A^* A^*)^{-1} \sigma^2 .\]

Now suppose that the unmodified model (4.1) is the true representation of the state of nature, that is, that \( \gamma_{a+t-2} \neq \gamma_{a+t-1} \). Denote the reparameterized model as:

\[ E(\mathbf{y}) = A^{*\prime}_{\mathbf{1}}^* \mathbf{y} \]
where
\[ A^*_{\tau} = [A^* \mid a] \]
\[ a' = [0', 1] \]
\[ A^*_{\tau} = \begin{pmatrix} I^* \\ \gamma_{a+t-1} - \gamma_{a+t-2} \end{pmatrix} . \]

Then (4.5) should be replaced by:
\[
E(\hat{\mu}) = (A^* A^*)^{-1} A^* A^*_{\tau} \hat{\mu} \\
= [I \mid (A^* A^*)^{-1} A^* a] \begin{pmatrix} I^* \\ \gamma_{a+t-1} - \gamma_{a+t-2} \end{pmatrix} \\
= I^* + (A^* A^*)^{-1} A^* a (\gamma_{a+t-1} - \gamma_{a+t-2}) .
\]

Thus it is seen that the estimates of the contrasts cannot all be unbiased estimates of the contrasts of interest. Some of the estimated contrasts are aliased by the contrast \( \gamma_{a+t-1} - \gamma_{a+t-2} \). This alias is not necessarily constant for all the estimates. Furthermore, we cannot correct for these aliases, since as we have seen, the contrasts of interest are not all estimable. If the effects of these two cohorts are nearly equal, however, the alias will be small.

As the second alternative model, suppose that a true representation of the state of nature is given by \( \gamma_{a+t-1} = \gamma_{a+t-2} \) and \( \gamma_{a+t-3} = \gamma_{a+t-2} \). Denote this model:

\[ (4.6) \quad E(\mu) = A^*_{\tau} \]

where \( A^*_{\tau} \) is obtained from \( A^* \) of (4.3) by combining the last two columns,
and \( \mathbf{T}_0^* \) is a comparable reduction of \( \mathbf{T}^* \) with \( \gamma_{m-1}^* \) deleted. In this case,

\[
\mathbf{E}(\mathbf{T}^*) = (\mathbf{A}^*\mathbf{A}^*)^{-1} \mathbf{A}^*\mathbf{A}^*\mathbf{T}_0^* \\
= \begin{bmatrix} \mathbf{I} & \mathbf{T}_0^* \\ \mathbf{0}^T \end{bmatrix}
\]

This demonstrates that none of the estimates are biased. They are not the unique minimum variance unbiased estimates under model (4.6), however.

These two examples indicate the effect on the estimates of an inappropriate model in the symmetric case. Intuitively, one would expect the biases to increase as the number of unequal cohort effects that are combined increases. The same reasoning holds for the non-symmetric case, where the bias may be expected to enter in a more complex way.

On the other hand, because the coefficient matrix determines the variances of the estimates, minimum variance is attained when the maximum number of cohorts are combined appropriately.

An added consideration is that the biased estimates may have greater precision than unbiased estimates, and may be more desirable if the bias is sufficiently small. It is difficult to find the coefficient matrix which would give the greatest precision because of the unbalanced nature of the data. Although, with no missing observations, each age interval has \( t \) observations, and each time interval has \( a \) observations, the number of observations in each cohort interval varies
from 1 to min(a, t) in the symmetric case. Thus two cohorts, depending
on how close in time they are, vary in the number of age and/or time
intervals they have in common. Cohorts which are widely separated may
have no common age or time intervals. The way in which cohorts are
combined can decrease this imbalance and increase precision.

The final consideration is the computational problem. There is no
simple general solution to the normal equations, due to the unbalanced
nature of the data. The size of the matrix to be inverted in most
practical problems makes inversion by desk calculator difficult. The
electronic computer also presents problems. At best, it can give only
approximate inverses. Multiple-precision programs required for the
more ill-conditioned matrices are not generally available, and there
has not yet been enough numerical analysis on the programs available
for inversion to estimate the limits of error involved.
CHAPTER V
EXAMPLES

Data Used

Because the cohort model is expected to be of most use in the interpretation of rates, two sets of data in which rates are involved are examined.

The data for study I, taken from Greenberg, et al. [3], represents the symmetric case, and are shown in Table 5.1. In this study \( a = 15 \) and \( t = 7 \) so that, using the definition of cohort intervals given for model (4.1), there are twenty-one one-year cohort intervals.

Table 5.1. Age specific discovery rates of primary and secondary syphilis among negro females per 1,000 population, 1941-1947, by age and year of discovery in study area

<table>
<thead>
<tr>
<th>Age</th>
<th>1941</th>
<th>1942</th>
<th>1943</th>
<th>1944</th>
<th>1945</th>
<th>1946</th>
<th>1947</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1.6</td>
<td>5.0</td>
<td>10.2</td>
<td>8.5</td>
<td>1.7</td>
<td>24.4</td>
<td>8.2</td>
</tr>
<tr>
<td>16</td>
<td>9.4</td>
<td>12.9</td>
<td>16.6</td>
<td>15.2</td>
<td>15.5</td>
<td>22.6</td>
<td>6.4</td>
</tr>
<tr>
<td>17</td>
<td>14.0</td>
<td>16.0</td>
<td>21.5</td>
<td>16.8</td>
<td>20.5</td>
<td>27.1</td>
<td>4.7</td>
</tr>
<tr>
<td>18</td>
<td>10.3</td>
<td>15.1</td>
<td>26.8</td>
<td>36.2</td>
<td>23.5</td>
<td>34.8</td>
<td>17.9</td>
</tr>
<tr>
<td>19</td>
<td>20.6</td>
<td>19.6</td>
<td>15.7</td>
<td>23.0</td>
<td>16.8</td>
<td>50.0</td>
<td>13.3</td>
</tr>
<tr>
<td>20</td>
<td>7.4</td>
<td>18.0</td>
<td>21.8</td>
<td>32.8</td>
<td>20.8</td>
<td>42.0</td>
<td>15.5</td>
</tr>
<tr>
<td>21</td>
<td>15.1</td>
<td>4.9</td>
<td>11.1</td>
<td>26.7</td>
<td>21.0</td>
<td>20.2</td>
<td>18.7</td>
</tr>
<tr>
<td>22</td>
<td>5.5</td>
<td>10.4</td>
<td>15.3</td>
<td>29.4</td>
<td>7.8</td>
<td>17.2</td>
<td>11.0</td>
</tr>
<tr>
<td>23</td>
<td>7.0</td>
<td>5.7</td>
<td>7.3</td>
<td>21.7</td>
<td>8.0</td>
<td>21.2</td>
<td>6.8</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>7.3</td>
<td>2.0</td>
<td>11.6</td>
<td>7.4</td>
<td>9.1</td>
<td>3.5</td>
</tr>
</tbody>
</table>

| 25  | 6.7  | 1.7  | 7.3  | 1.9  | 1.8  | 8.0  | 5.1  |
| 26  | 5.1  | 1.8  | 7.4  | 1.9  | 2.1  | 8.0  | 5.0  |
| 27  | 3.7  | 3.5  | 3.7  | 3.8  | 4.0  | 5.8  | 5.2  |
| 28  | 1.8  | 1.9  | 0    | 5.7  | 0    | 1.8  | 3.8  |
| 29  | 1.8  | 5.5  | 1.9  | 0    | 3.8  | 12.4 | 0    |

\(^a\) The ages reported were to nearest birthday. The ages in this table, therefore, represent the midpoints of yearly intervals.
The data for study II were given in a paper by MacMahon [9] and later analyzed by Hussein [6]. They are shown in Table 5.2. This study represents the non-symmetric case, involving sixty-one one-year cohort intervals.

Table 5.2. Age-specific "incidence" rates for breast cancer in females per 100,000 population, Connecticut, 1935-1951

<table>
<thead>
<tr>
<th>Age in years</th>
<th>1935-1937</th>
<th>1938-1942</th>
<th>1943-1947</th>
<th>1948-1951</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-34</td>
<td>13.2</td>
<td>18.6</td>
<td>16.1</td>
<td>21.6</td>
</tr>
<tr>
<td>35-39</td>
<td>37.0</td>
<td>37.4</td>
<td>41.4</td>
<td>49.0</td>
</tr>
<tr>
<td>40-44</td>
<td>72.4</td>
<td>84.6</td>
<td>76.1</td>
<td>82.7</td>
</tr>
<tr>
<td>45-49</td>
<td>94.6</td>
<td>88.2</td>
<td>105.3</td>
<td>115.5</td>
</tr>
<tr>
<td>50-54</td>
<td>114.1</td>
<td>109.3</td>
<td>100.4</td>
<td>125.2</td>
</tr>
<tr>
<td>55-59</td>
<td>179.0</td>
<td>160.5</td>
<td>140.9</td>
<td>144.9</td>
</tr>
<tr>
<td>60-64</td>
<td>178.0</td>
<td>171.6</td>
<td>157.3</td>
<td>160.1</td>
</tr>
<tr>
<td>65-69</td>
<td>201.2</td>
<td>209.2</td>
<td>183.3</td>
<td>186.8</td>
</tr>
<tr>
<td>70-74</td>
<td>192.1</td>
<td>235.9</td>
<td>237.5</td>
<td>235.7</td>
</tr>
</tbody>
</table>

Methods of Analysis

Three general methods of modification were used in these studies to obtain models which permitted all contrasts of interest to be estimated. The first method is that of simply defining new cohort intervals by combining adjacent basic cohorts. Abrupt changes would not generally be expected in consecutive cohort effects, so it was anticipated that combining adjacent cohorts would produce negligible if any biases. For study I, as discussed earlier in connection with model (4.1), at least one such combination is necessary in order that all contrasts of interest
may be estimated. More combinations may be desirable in order to increase the precision of the estimates.

For study II, enough basic cohorts must be combined so that not more than twelve new cohorts are defined. The coefficients of each new cohort effect contributing to an observation is determined as indicated in model (1.3). That is, the coefficient is proportional to the contribution of its cohort effect to the observation, and such that the cohort coefficients for any observation sum to one.

The second method of modifying the basic model was suggested by Case's [1] work. If, in the symmetric case, basic cohort intervals are defined to be $1/n$ of the basic time unit in width, and then these basic cohorts are recombined so that only two cohort effects are associated with each observation, then the coefficients of the two effects are:

\[
\frac{n(n - 1)}{2n^2} = \frac{1}{2} \left( 1 - \frac{1}{n} \right)
\]

and

\[
\frac{n(n + 1)}{2n^2} = \frac{1}{2} \left( 1 + \frac{1}{n} \right).
\]

As $n$ tends to infinity, clearly both coefficients tend to $1/2$. In geometric representation, this implies a straight line bisecting the area for a rate, as follows:

![Diagram](image-url)
The model then becomes:

\[ E(y_{ij}) = \mu + \alpha_i + \beta_j + \frac{1}{2} \gamma_k + \frac{1}{2} \gamma_{k+1} \]

where \( k = i + j - 1 \). This is similar to taking moving averages of the basic cohort effects.

The rank of \( V^+(A) \) for this model is 5, but can be reduced to the required 3 in a manner similar to that of the first method, that is, by combining cohorts. For study I, the maximum number of cohorts that can be defined if all contrasts of interest are to be estimable is twenty, the same as by the first method. The extension of this second method to the non-symmetric case is seen by drawing straight lines over the geometric representation of the data. The coefficients of the cohorts are determined in a manner analogous to that for the first method.

The third method of modifying the basic model is to assume that the relation cohort effects to time (i.e. to the time of origin of each cohort) follows a polynomial function. If such a function is appropriate, it is desirable to fit a polynomial of degree high enough to disclose all the real effects, but not so high as to allow "noise" of innate variability to obscure real effects. This method also would tend to keep the number of parameters estimated smaller and thus increase the precision of the estimates. \( R^+(A) \) for the model using this method must be 2, if all contrasts of interest are to be estimable, with \( V^+(A) = V(G) \) where

\[
G (2 \times p) = \begin{bmatrix}
-1 & 1^t & 0^t & 0^t \\
-1 & 0^t & 1^t & 0^t
\end{bmatrix}
\]
For study I, fitting a polynomial including a linear term to the twenty-one one-year cohorts makes $R^+(A) = 3$. The above condition can be met by combining two of the cohorts. To retain symmetry, cohorts 1 and 2 and cohorts 20 and 21 were combined for modifications using this method. An additional modification using this method for study I assumes an underlying model for age, that is, that there is a polynomial function of age appropriate to describe the age effect. With this modification, $V^+(A) = V(C)$ where

$$G(1 \times p) = [-1, 0_1', 0_2', \ldots, 0_p']$$

To use this method for study II it is necessary to find, for each observation, the coefficient of the regression parameters as a linear combination of coefficients appropriate for the basic cohorts. The linear combination is determined so that each basic cohort is weighted in proportion to its contribution to the observation. For example, the cohorts contributing to $y_{11}$ are 1 to 7 as follows:

<table>
<thead>
<tr>
<th>1955 - 1997</th>
<th>5 6 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 5 6</td>
<td></td>
</tr>
<tr>
<td>3 4 5</td>
<td></td>
</tr>
<tr>
<td>2 3 4</td>
<td></td>
</tr>
<tr>
<td>1 2 3</td>
<td></td>
</tr>
</tbody>
</table>

age 70-74

The coefficient for the $r^{th}$ regression parameter is $z_{qr}(r)$, and for this particular case
\[ z' = (1/15, 2/15, 3/15, 3/15, 3/15, 2/15, 1/15) \]

and

\[ z^{(r)*} = (1^r, 2^r, 3^r, 4^r, 5^r, 6^r, 7^r). \]

If a logarithmic term is included in the function, which is otherwise a polynomial, then the same linear combination is taken of the logarithms of the elements of the cohort vector. For the modifications of study II using this method, \( v^+(A) = v(0) \) where

\[
G (2 \times p) = \begin{pmatrix}
-1 & 1^i_1 & 0^f_2 & 0^f_3 \\
-1 & 0^f_1 & 1^i_2 & 0^f_3
\end{pmatrix}
\]

as required.

Results of Study I

For this study, the transformation \( y_{ij} = \log e (r_{ij} + 1) \) is used for all modifications. The specific modifications used for the first two methods are presented in Table 5.3 in terms of the twenty-one basic one-year cohorts defined for model (4.1); for the third method Table 5.3 indicates the polynomial function employed, the regression being on the 19 cohorts remaining after combining cohorts 1 and 2, and 20 and 21. The partitions \( A_0, A_1, \) and \( A_2 \) of the coefficient matrix are given by (3.1) for all modifications except 4-1 and 4-2. For these two modifications \( A_1 \) is replaced by the matrix appropriate for the polynomial function that is employed to represent the underlying age function.

The estimates obtained by the various modifications are presented in Figures 5.1 through 5.11. The estimates of the contrasts are
Table 5.3. Table to show which basic cohorts are combined in each particular modification of the model - study 2

<table>
<thead>
<tr>
<th>Modification</th>
<th>No. of cohorts</th>
<th>Cohorts of the modification</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td></td>
<td></td>
<td>1-1</td>
<td>10</td>
<td>1-2</td>
<td>3-4</td>
<td>5-6</td>
<td>7-8</td>
<td>9-10</td>
<td>11-12</td>
<td>13-14</td>
<td>15-16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1-2</td>
<td>7</td>
<td>1-3</td>
<td>4-6</td>
<td>7-9</td>
<td>10-12</td>
<td>13-15</td>
<td>16-18</td>
<td>19-21</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1-3</td>
<td>7</td>
<td>1-5</td>
<td>6-7</td>
<td>8-9</td>
<td>10-11</td>
<td>12-13</td>
<td>14-15</td>
<td>16-21</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1-4</td>
<td>5</td>
<td>1-4</td>
<td>5-8</td>
<td>9-12</td>
<td>13-16</td>
<td>17-21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td></td>
<td></td>
<td>2-1</td>
<td>10</td>
<td>1-3</td>
<td>3-5</td>
<td>5-7</td>
<td>7-9</td>
<td>9-11</td>
<td>11-13</td>
<td>13-15</td>
<td>15-17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2-2</td>
<td>5</td>
<td>1-5</td>
<td>5-9</td>
<td>9-13</td>
<td>13-17</td>
<td>17-21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method 3</td>
<td>Function fitted to cohorts; basic cohorts 1 and 2, 20 and 21 combined</td>
<td></td>
<td>3-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cohort function includes linear term</td>
<td></td>
<td>3-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cohort function includes logarithmic term and 4th degree polynomial</td>
<td></td>
<td>4-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Function fitted to 15 age intervals also</td>
<td></td>
<td>4-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2/ The basic cohorts are numbered from 1 to 21. It will be noted that in method 2, some basic cohorts occur in two adjacent cohorts of the modified model; in these cases half of the basic cohort is apportioned to each of the adjacent cohorts of the modified model.
Figure 5.1. Estimated relative age effects for study I, modifications 1-1 through 1-4.
Figure 5.2. Estimated relative time effects for study I, modifications 1-1 through 1-4

Figure 5.3. Estimated relative cohort effects for study I, modifications 1-1 through 1-4
Figure 5.4. Estimated relative age effects for study I, modifications 2-1 and 2-2
Figure 5.5. Estimated relative time effects for study I, modifications 2-1 and 2-2
Figure 5.6. Estimated relative cohort effects for study I, modifications 2-1 and 2-2
Figure 5.7. Estimated relative age effects for study I, modifications 3-1 and 3-2
Figure 5.8. Estimated relative time effects for study I, modifications 3-1 and 3-2
Figure 5.9. Estimated regression on cohorts for study I, modifications 3-1, 3-2, and 4-1.
Figure 5.10. Estimated regression on ages for study I, modifications 4-1 and 4-2
Figure 5.11. Estimated relative time effects for study I, modifications 4-1 and 4-2
obtained by a reparameterization appropriate to the modification used, in the same manner as in model (4.3). Where a polynomial function represents the effect of a factor in a modification, that is, in modifications 3-1, 3-2, 4-1, and 4-2, there is no reparameterization of the regression coefficients for that factor. For ease in comparing results however, the estimates of each set of contrasts are presented in such a way that the sum of the effects for any one factor is zero. For example, for time effects:

$$\hat{\beta}_t = -\frac{1}{t} \sum_{j=1}^{t-1} \hat{\beta}_j^*$$

and

$$\hat{\beta}_j = \hat{\beta}_j^* - \hat{\beta}_t^*, \quad j = 1, 2, ..., t - 1.$$  

The line segments drawn through the consecutive estimates of the age effects in Figures 5.1, 5.4, 5.7, and 5.10 have somewhat similar conformations for all modifications. The modifications 2-1, 3-2, and 4-1 show the most extreme departure. They are similar to the others, but have the appearance of having been stretched and rotated clockwise.

The same remarks can be made for the line segments through the estimated time effects in Figures 5.2, 5.5, 5.8, and 5.11 except that here the line segments for the modifications 2-1, 3-2, and 4-1 appear to be rotated counterclock-wise.

In the case of the cohort effects the various modifications also appear to have the effect of rotating the line segments, as is seen in Figures 5.3, 5.6, and 5.9, with the modifications 2-1, 3-2, and 4-1 rotated similarly to their age counterparts. For this factor, however,
the particular modification alters the inferences that would be made quite a bit. A decreasing trend in rates for more recent cohorts is indicated by modifications 1-2 and 3-1, as well as 2-1, 3-2, and 4-1, while the other modifications indicate an increasing trend.

One way of judging the computing error due to continual truncation and/or the bias involved in the estimates is to compare estimates of the mean from different modifications. Under the assumption that a properly weighted sum of the effects for each factor is zero, one estimate of the mean is given by

\[ \hat{\mu}_1 = \sum \sum y_{ij} / n. \]

For the first two methods, a second estimate is given by:

\[ \hat{\mu} = \hat{\mu}^* - \hat{\alpha}_a - \hat{\beta}_t - \hat{\gamma}_k, \]

where \( \hat{\alpha}_a, \hat{\beta}_t, \) and \( \hat{\gamma}_k \) are found as described above. For modifications 3-1, and 3-2 of the third method, under the assumption that the sum of age effects and time effects is zero, \( \mu \) is

\[ \text{E} \left( \sum \sum y_{ij} / n \right) = \mu^* - \alpha_a - \beta_t + \sum \sum \sum \sum z^r / n. \]

where \( \theta_r \) is the \( r^{th} \) regression coefficient of the polynomial function of the cohorts. Similarly for modifications 4-1 and 4-2, assuming the sum of time effects is zero, an estimate of the mean is

\[ \text{E} \left( \sum \sum y_{ij} / n \right) = \mu^* - \beta_t + \sum \lambda_s \sum \sum w^s / n + \sum \theta_r \sum \sum z^r / n. \]

where \( \lambda_s \) is the \( s^{th} \) regression coefficient of the polynomial function of age, and \( \theta_r \) is the \( r^{th} \) regression coefficient of the polynomial.
function of cohorts (all zero for modification 4-2). For the third method, estimates of the mean are found by replacing the above parameters by their estimates. These estimates are presented in Table 5.4, together with the mean square error for each modification. The error sums of squares were obtained as differences between two sums of squares by the same computer program after the estimates for each modification were computed.

Results of Study II

The transformation $y_{ij} = \log_e (r_{ij})$ is used for all modifications of this study. The specific modifications for this study are found in Table 5.5 in terms of the sixty-one basic one-year cohorts for the first two methods, and the polynomial function for the third method. The partitions $A_0$, $A_1$, and $A_2$ of the coefficient matrix are given by (3.1) for all modifications.

The estimates are obtained for the modifications in the manner described for study I. The estimated relative age effects are presented in Figures 5.12, 5.15, and 5.18. Figures 5.13, 5.16, and 5.19 illustrate the estimated relative time effects. The estimated relative effects for cohorts are shown in Figures 5.14 and 5.17, and the estimated regression on cohorts in Figure 5.20.

In this study, there is some similarity in the line segments drawn through the estimated effects for each factor. The amount of expansion and rotation of a cohort line segment for any particular modification appears to correspond to that of the age line segment for that modification with the same direction of rotation and to that of the time line
<table>
<thead>
<tr>
<th>Modification</th>
<th>( \hat{\mu} )</th>
<th>M S E</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>2.35</td>
<td>.25657</td>
<td>75</td>
</tr>
<tr>
<td>1-2</td>
<td>2.14</td>
<td>.25035</td>
<td>78</td>
</tr>
<tr>
<td>1-3</td>
<td>2.28</td>
<td>.26940</td>
<td>78</td>
</tr>
<tr>
<td>1-4</td>
<td>2.48</td>
<td>.25446</td>
<td>80</td>
</tr>
<tr>
<td>2-1</td>
<td>-3.13</td>
<td>.25202</td>
<td>75</td>
</tr>
<tr>
<td>2-2</td>
<td>2.50</td>
<td>.25703</td>
<td>80</td>
</tr>
<tr>
<td>3-1</td>
<td>3.86</td>
<td>.26352</td>
<td>83</td>
</tr>
<tr>
<td>3-2</td>
<td>11.41</td>
<td>.24507</td>
<td>79</td>
</tr>
<tr>
<td>4-1</td>
<td>1.60</td>
<td>.25163</td>
<td>90</td>
</tr>
<tr>
<td>4-2</td>
<td>1.23</td>
<td>.25749</td>
<td>94</td>
</tr>
<tr>
<td>( \hat{\mu} )</td>
<td>2.189</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.5. Table to show which basic cohorts are combined in each particular modification of the model - study II a/

<table>
<thead>
<tr>
<th>Modification</th>
<th>No. of cohorts</th>
<th>Cohorts of the modification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method 1</strong></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1-1</td>
<td>11</td>
<td>1-6</td>
</tr>
<tr>
<td>1-2</td>
<td>8</td>
<td>1-5</td>
</tr>
<tr>
<td>1-3</td>
<td>8</td>
<td>1-13</td>
</tr>
<tr>
<td>1-4</td>
<td>7</td>
<td>1-9</td>
</tr>
<tr>
<td>1-5</td>
<td>7</td>
<td>1-15</td>
</tr>
<tr>
<td>1-6</td>
<td>6</td>
<td>1-16</td>
</tr>
<tr>
<td>1-7</td>
<td>6</td>
<td>1-11</td>
</tr>
<tr>
<td>1-8</td>
<td>5</td>
<td>1-13</td>
</tr>
<tr>
<td>1-9</td>
<td>4</td>
<td>1-16</td>
</tr>
<tr>
<td><strong>Method 2</strong></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>2-1</td>
<td>11</td>
<td>1-7</td>
</tr>
<tr>
<td><strong>Method 3</strong></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>3-1</td>
<td></td>
<td>cohort function includes logarithmic term and 3rd degree polynomial</td>
</tr>
<tr>
<td>3-2</td>
<td></td>
<td>cohort function includes 4th degree polynomial</td>
</tr>
<tr>
<td>3-3</td>
<td></td>
<td>cohort function includes logarithmic term and 4th degree polynomial</td>
</tr>
</tbody>
</table>

a/ The basic cohorts are numbered from 1 to 61. It will be noted that in method 2 some basic cohorts occur in two adjacent cohorts of the modified model; in these cases half of the basic cohort is apportioned to each of the adjacent cohorts of the modified model.
Figure 5.12. Estimated relative age effects for study II, modifications 1-1 through 1-9

Figure 5.13. Estimated relative time effects for study II, modifications 1-1 through 1-9
Figure 5.14. Estimated relative cohort effects for study II, modifications 1-1 through 1-9
Figure 5.15. Estimated relative age effects for study II, modifications 2-1 and 2-2

Figure 5.16. Estimated relative time effects for study II, modifications 2-1 and 2-2
Figure 5.17. Estimated relative cohort effects for study II, modifications 2-1 and 2-2
Figure 5.18. Estimated relative age effects for study II, modifications 3-1, 3-2, and 3-3

Figure 5.19. Estimated relative time effects for study II, modifications 3-1, 3-2, and 3-3
Figure 5.20. Estimated regression on cohorts for study II, modifications 3-1, 3-2, and 3-3
segment with the opposite direction of rotation. In this study, inferences about the trends for time effects, effects of the older age groups, and cohort effects are very much affected by the particular modification used.

The estimates of the mean, found as for study I, and mean square errors are presented in Table 5.6. The error sum of squares for modification 1-1 was obtained as the sum of squares of deviations of the observed from predicted values, because the value given by the computer program (which obtained a difference between two sums of squares) was negative. The error sums of squares for modifications 3-1, 3-2, and 3-3 were obtained by the formula:

\[ \text{SSE} = \mathbf{Y}' \mathbf{Y} - \mathbf{1}' \mathbf{A}' \mathbf{Y}, \]

because the program used to obtain the estimates for these modifications did not provide the mean square error. For the other modifications, the mean square error was given by the computer program.

It is difficult to obtain good inverses of the matrix \((A^* A^*)\) for modifications 3-1, 3-2, and 3-3. Although it is true that the matrices to be inverted are non-singular, they are ill-conditioned. For this reason, a double precision matrix inversion computer program was used to obtain estimates for these modifications. Even so, for modification 3-2 the element \((1,1)\) of the inverse was negative. This illustrates the mechanical difficulties that can be encountered in obtaining solutions for any general linear hypothesis.
Table 5.6. Estimates of the mean and mean square error - study II

<table>
<thead>
<tr>
<th>Modification</th>
<th>$\hat{\mu}$</th>
<th>MSE</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>2.68</td>
<td>.00417</td>
<td>14</td>
</tr>
<tr>
<td>1-2</td>
<td>1.93</td>
<td>.00530</td>
<td>17</td>
</tr>
<tr>
<td>1-3</td>
<td>2.35</td>
<td>.00532</td>
<td>17</td>
</tr>
<tr>
<td>1-4</td>
<td>2.72</td>
<td>.00497</td>
<td>18</td>
</tr>
<tr>
<td>1-5</td>
<td>2.58</td>
<td>.00341</td>
<td>18</td>
</tr>
<tr>
<td>1-6</td>
<td>2.22</td>
<td>.00860</td>
<td>19</td>
</tr>
<tr>
<td>1-7</td>
<td>1.89</td>
<td>.00678</td>
<td>19</td>
</tr>
<tr>
<td>1-8</td>
<td>2.29</td>
<td>.00783</td>
<td>20</td>
</tr>
<tr>
<td>1-9</td>
<td>2.09</td>
<td>.00904</td>
<td>21</td>
</tr>
<tr>
<td>2-1</td>
<td>-1.75</td>
<td>.00191</td>
<td>14</td>
</tr>
<tr>
<td>2-2</td>
<td>2.81</td>
<td>.00427</td>
<td>16</td>
</tr>
<tr>
<td>3-1</td>
<td>2.83</td>
<td>11.74967</td>
<td>20</td>
</tr>
<tr>
<td>3-2</td>
<td>2.71</td>
<td>12.38399</td>
<td>20</td>
</tr>
<tr>
<td>3-3</td>
<td>-0.05</td>
<td>.08318</td>
<td>19</td>
</tr>
<tr>
<td>$\hat{\mu}$</td>
<td>4.566</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion and Conclusions

It is difficult to draw unequivocal conclusions from the results obtained. The issue is clouded by the computational errors associated with manipulation of ill-conditioned matrices. One indication of the difficulty is the poor agreement between $\hat{\mu}$ and $\mu_1$ in many instances, although bias may contribute to this lack of agreement. It must be recognized, however, that even good agreement does not guarantee that no computational errors or biases exist in the estimates.

Another indication of computational errors is afforded by the check made on the inverses incorporated in the computer program employed to obtain solutions for the first two methods of modification. This check, made by premultiplying the original matrix by its inverse, should yield an identity matrix. There is considerable variation in these checks. The off-diagonal element of largest absolute magnitude is 0.0188; it occurs in modification 2-1 of study II. That is the modification for which the estimated mean shows the greatest deviation from $\hat{\mu}_1$ for study II. This method of checking the inverse, although giving an indication of a poor inverse, does not provide a measure of how bad it may be. And again, a good check does not insures a good inverse.

Even if we assume insignificant computational errors, there is still lack of a universally acceptable criterion for determining the adequacy of a model. One criterion, suggested by Webster [12] in considering whether or not to include an additional term in a polynomial regression, is to use the model that yields the minimum "mean square error." The "mean square error" contains deviations attributable to
differences between the true and the assumed model as well as the deviations caused by random fluctuations from the true model. One inherent difficulty in using this criterion is the problem of knowing when the minimum mean square error has been attained, since the number of modifications possible are not denumerable.

A second criterion that could be used is to choose the model for which the estimates have the smallest variances. For the first method of modification, the precision generally increases as the number of cohorts defined in the modification and thus the number of parameters in the model, decreases. Also, the precision tends to increase as the numbers of observations associated with the several cohorts tend to equality. The reduction of the number of parameters by the use of a smooth function rather than a step function to represent the cohort effects did not necessarily, however, increase the precision as desired.

It is not clear how much reliance can be placed on the values obtained for the mean square error and for the estimated variances of the estimates because of the computational errors.

The general impression gained from these results is not, however, indistinct. With the modification methods 1 and 2 the estimates tend to stability as the number of cohorts defined decreases. It seems that such stable patterns must reflect meaningful trends in the data. If it were possible to split the data so that more than one set of concurrent data could be analyzed, one could ascertain more clearly whether the estimates would tend to agree more as the number of cohorts is decreased to some optimal point.
The results from the smooth function representation were not as promising as hoped, but they suggest some value to the smooth function approach. In study I, particularly, the modifications with more parameters show more fluctuations in the estimates than those with fewer parameters. One point that should be made is that the estimates of the parameters are inherently highly correlated because the cohorts are part of the age-time interaction. This is recognized from the fact that in study I, not all 21 cohorts could be used to estimate the linear term of the cohort function, and in study II, although all 61 cohorts could be used to fit the linear cohort term, the matrices were highly ill-conditioned.

It is recognized that the modifications studied here far from exhaust the many ways of modifying that might lead to useful estimates. It is clear, however, that a large class of modifications are of types not worth studying. The results suggest a certain class of modifications that could be effective. This class will be noted under suggestions for future work.
CHAPTER VI
SUMMARY AND RECOMMENDATIONS

Resume of Results

Problems of analyzing rate data were studied in the context of the following model:

\[ E(y_{ij}) = \mu + \sum_{u} w_{uij} \alpha_{u}^* + \sum_{v} x_{vij} \beta_{v}^* + \sum_{g} z_{gij} \gamma_{g}^* \]

where

- \( y_{ij} \) = \( \log (r_{ij} + h) \); \( r_{ij} \) = the rate given for the \( i^{th} \) age interval and the \( j^{th} \) time period; \( r_{ij} + h > 0 \)
- \( \mu \) = general mean
- \( \alpha_{u} \) = the \( u^{th} \) age effect in the \( i^{th} \) age interval
- \( \beta_{v} \) = the \( v^{th} \) time effect in the \( j^{th} \) time period
- \( \gamma_{g} \) = the \( g^{th} \) cohort effect in the \( i^{th} \) age interval and the \( j^{th} \) time period
- \( w_{uij} \) = the coefficient for the \( u^{th} \) age effect determined by its contribution to the \( ij^{th} \) rate
- \( x_{vij} \) = the coefficient of the \( v^{th} \) time effect determined by its contribution to the \( ij^{th} \) rate
- \( z_{gij} \) = the coefficient of the \( g^{th} \) cohort effect determined by its contribution to the \( ij^{th} \) rate.

For the model to be reasonable, it must be that

\[ \sum_{u} w_{uij} = \sum_{v} x_{vij} = \sum_{g} z_{gij} = 1 \]

This basic step function model can be altered to allow for other ways
of representing the effects of a factor. For example, if the cohort factor is to be represented by a polynomial, $\sum g_{ij} \gamma_g$ is replaced by $\sum z^{*}_{ij} \gamma_r$ where $z^{*}_{ij} = \sum g^T z_{ij}$ and $\gamma_r$ is the regression coefficient in the polynomial for the $r$th degree term.

With this basic model (6.1), it is desirable to estimate all contrasts of the type $\sum c_g \gamma_g$ for any factor in order to be able to make useful interpretations. A theorem, which for this model states the necessary and sufficient condition for all the desired quantities to be estimable is proved. The condition is that the rank of the coefficient matrix must be $p - 3$, where $p$ is the number of parameters in the model. For modifications of the model in which the step function representation of the effects of a factor is replaced by a smooth function representation, the condition is that the rank of the coefficient matrix must be $p - r$, where $r$ is the number of factors represented by step function effects.

In general, it was shown that, for model (6.1) or a modification of it, all of the useful estimates cannot be made without accepting additional assumptions. For example, in the simple special case where all time, age, and cohort intervals are of equal length, the model can be written as

$$(6.2) \quad \beta_i = \mu + \alpha_i + \beta_j + \gamma_k.$$  

The coefficient matrix for this model has a rank less than $p - 3$. To make the rank equal to $p - 3$, $p$ must be reduced. This can be done, for example, by assuming that $\gamma_k = \gamma_{k+1}$ for any $k < m$.  

Making assumptions of the sort just noted in order to obtain estimability can lead to estimable contrasts which are not unbiased estimates of the contrasts of interest. If, however, the assumptions made to obtain estimability do in fact hold, then, of course, unbiased estimates of the contrasts of interest are obtained.

It may be that assumptions, which simplify the model even further than do those required for estimability, actually hold. In this event, if the stronger assumptions are not made, the estimates of contrasts are all unbiased but are not minimum variance. Thus, one would wish to employ a model which is as simple as is compatible with the situation under study and which satisfies also the estimability condition.

Data from two sources were studied. Initially the number of cohorts defined was the maximum which permitted all contrasts of interest to be estimated. The basic cohorts were defined in two different ways. Under model (6.1), the number of cohorts was 20 in one study and 12 in the other. However, the matrices involved were so ill-conditioned that estimates could not be obtained because of computational problems.

Modifications were made in the model by further combination or collapsing of the basic cohorts. The minimum number defined was 5 in one study and 4 in the other. For modifications involving the larger numbers of cohorts, the matrices involved were always ill-conditioned, leading to large computational errors and high variances for the estimates. As the numbers were reduced, the computational difficulties diminished and the variances became smaller.
At an intermediate point in the reduction of the number of cohorts, the trends took a pattern which changed only moderately as the number of cohorts was further reduced. Results were the same for both ways of defining the basic cohorts. These stable trends were regarded as reflecting meaningful trends in the data.

A smooth function representation of the cohorts as a means of modification was also employed. In all cases the maximum number of cohorts compatible with estimability was defined. This produced very ill-conditioned matrices and the results were not encouraging.

Recommendations

Until further studies of this problem are done, only tentative recommendations can be made, and they cannot be very specific. From the results obtained it is recommended that the step function model (6.1) or (6.2) be used, starting with a large but estimable number of cohort contrasts, and then, in successive analyses, reducing the number of cohorts until further reduction does not materially change the pattern of trends. In the process of reduction it is only reasonable to combine adjacent cohorts, for this would be expected to keep biases at a minimum. Just what rules to follow in the reduction process needs further study. It is suggested, however, that combining of cohorts should be done only in the regions where the trend changes are minor.

It is suggested that, whenever possible, concurrent sets of data be collected in order that the several sets may be independently analyzed. Agreement and disagreement of results between sets when analyzed under various model modifications should help distinguish meaningful from spurious trends.
The smooth function approach has appeal on the grounds that it offers the possibility of faithfully reflecting cohort trends with a minimum of parameters. Results with this approach were, however, not encouraging. It is recommended, therefore, that the smooth function representation not be used until it has been studied further.

Suggestions for Future Work

The computational problems encountered are important but not unique to this problem. This leads to the suggestion that additional attention be given to the evaluation of computation errors so that results can be compared with more confidence.

The promise of revealing meaningful trends shown in the step function approach leads to the suggestion that an investigation be made for a criterion for determining the optimum number and allocation of the cohorts.

The smooth function approach, although intuitively appealing in that it reduces the number of parameters, did not produce satisfactory results as treated here. This was largely due to the high correlations between the parameters, particularly those of the linear and possibly quadratic terms of the cohort effects with the age and time effects. Reducing the number of cohorts defined results in diminishing the troublesome correlations. This fact leads to the suggestion that an effective approach might incorporate both a reduction in the number of cohorts defined in the model and a smooth function description of the cohort effects. It is interesting to note that in using this approach, the number of cohorts defined need not be the same for the several
regression coefficients used to describe the cohort effects. The author hopes to investigate this possibility in the near future.
LIST OF REFERENCES


331. Potthoff, Richard F. A test of whether two regression lines are parallel when the variances may be unequal. August, 1962.
334. Potthoff, Richard F. A test of whether two parallel regression lines are the same when the variances may be unequal. August, 1962.
343. Kendall, Peter. Estimation of the mean life of the exponential distribution from grouped data when the sample is censored—With application to life testing. February, 1963.
344. Roberts, Charles D. An asymptotically optimal sequential design for comparing several experimental categories with a standard or control. 1963.
356. Page, E. S. Controlling the standard deviation by cusums and warning lines.