THE MULTIPLE CLASSIFICATION MODEL IN AGE-PERIOD-COHORT ANALYSIS: THEORETICAL CONSIDERATIONS

by

Lawrence L. Kupper and Joseph M. Janis

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1311

November 1980

THE MULTIPLE CLASSIFICATION MODEL IN AGE-PERIOD-COHORT ANALYSIS: THEORETICAL CONSIDERATIONS

Lawrence L. Kupper and Joseph M. Janis

Department of Biostatistics
School of Public Health
University of North Carolina
Chapel Hill, NC 27514

SUMMARY

Age-period-cohort analysis has received considerable attention in the literature. If age and period (year) of observation are categorized to form a two-way table, the diagonals of the table represent categories of the third variable, birth cohort. The resulting multiple classification model containing the age, period, and cohort effects has an identification problem due to an inherent linear dependency among the variables. Previous authors have attempted to circumvent this problem by equating two or more effects based on subjective criteria, and then estimating the parameters via standard full-rank model procedures. It is shown in this paper that equating two or more effects does not pertain to the actual structure of the linear dependency, and that severe multicollinearity is the reason why previous work is inconclusive. Least squares estimation of the multiple classification model after reparametrization to eliminate the specific form of linear constraint present is recommended as a method for obtaining accurate estimates of the effects of interest. The equivalence of this estimation procedure to that of principal component regression analysis is also discussed.

Key Words: Age-period-cohort analysis; Multiple classification model; Multicollinearity; Reparametrization; Biased estimation; Principal component regression analysis.

1. INTRODUCTION

Cohort analysis (or, more specifically, age-period-cohort analysis) concerns methods for statistically analyzing data gathered on human populations followed over time, the purpose of such analysis being to quantify the separate effects of aging, (historical) period of time (e.g., as measured by year of death), and cohort membership (e.g., as measured by year of birth). Data of this type arise quite often in such diverse fields as sociology (Mason et al. [1973]); developmental psychology (Schaie [1965], Baltes [1968]); epidemiology (Greenberg et al. [1950], Hussein [1961], Barrett [1973]); and political science and economics (Knoke and Hout [1974], Winsborough [1975]). A careful examination of the pertinent literature reveals that the state-of-the-art with regard to the statistical treatment of such data is still in its early stages of development.

The most recent and detailed discussions of the analysis of age-period-cohort data have been given by Portman [1963], Mason et al. [1973], and Mason and Fienberg [1978] in the following context. If age (A) and period (P) of observation are categorized to form a two-way table, the diagonals of the table represent categories of the third variable, birth cohort (C). The multiple classification model containing the age, period, and cohort effects corresponding to this categorization has what is called an *identification problem*, which pertains to the presence of an inherent linear dependency among these three categorical variables. This identification problem is easily recognized if age,

period (e.g., year of death), and cohort (e.g., year of birth) are *continuous* variables; then, because $P = A + C$, it is impossible to estimate uniquely each of the parameters in the model $E(Y) = \mu + \alpha A + \beta P + \gamma C$. The authors cited above have attempted to circumvent this problem by equating two or more effects based on subjective criteria (e.g., a priori conjectures), and then estimating the parameters of the resulting artificially constrained multiple-classification model via standard full-rank regression procedures.

It will be shown in this paper that equating two or more effects does not pertain to the actual structure of the inherent linear dependency basic to the identifiability problem, and that severe multicollinearity is the major reason why previous model fitting efforts have led to unstable and inconclusive results. We will derive the heretofore unknown structure of this linear dependency in Section 3. This theoretical result will lead us to recommend a method of data analysis producing stable and accurate estimates of the effects of interest.

2. THE MULTIPLE CLASSIFICATION MODEL: GENERAL CONSIDERATIONS

Consider the multiple classification model

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{a-i+j} , \qquad (2.1)$$

where Y_{ij} is the observed response (e.g., a mean, a rate, a log rate, a logit, etc.) in the (i, j)-th cell in the cross-classification, α_i is the effect of the i-th age group, β_j is the effect of the j-th period, and γ_{a-i+j} is the cohort effect associated with the i-th age group and j-th period, $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, p$. Note that there are $(a + p - 1)$ cohort effects associated with model (2.1).

The special case $a = 3$, $p = 4$ is diagrammed below:

		PERIOD (j)			
		$j = 1$	$j = 2$	$j = 3$	$j = 4$
AGE(i)	$i = 1$	$\mu + \alpha_1 + \beta_1 + \gamma_3$	$\mu + \alpha_1 + \beta_2 + \gamma_4$	$\mu + \alpha_1 + \beta_3 + \gamma_5$	$\mu + \alpha_1 + \beta_4 + \gamma_6$
	$i = 2$	$\mu + \alpha_2 + \beta_1 + \gamma_2$	$\mu + \alpha_2 + \beta_2 + \gamma_3$	$\mu + \alpha_2 + \beta_3 + \gamma_4$	$\mu + \alpha_2 + \beta_4 + \gamma_5$
	$i = 3$	$\mu + \alpha_3 + \beta_1 + \gamma_1$	$\mu + \alpha_3 + \beta_2 + \gamma_2$	$\mu + \alpha_3 + \beta_3 + \gamma_3$	$\mu + \alpha_3 + \beta_4 + \gamma_4$

Portman [1963] and Mason et al. [1973] have shown that, in general, pairwise contrasts of interest like $(\alpha_i - \alpha_{i'})$, $(\beta_j - \beta_{j'})$, and $(\gamma_k - \gamma_{k'})$ cannot be estimated *unbiasedly* unless, for example, two age, period, or cohort effects can be assumed to be equal in model (2.1). However, as Searle [1971, p. 212] points out, assuming that two effects are equal for purposes of solving the normal equations only leads to unbiased estimates of such pairwise contrasts if that equality assumption about the parameters is actually true. Since, in reality, such an assumption is usually not valid, concern about obtaining unbiased estimates is unwarranted. More importantly, however, such constraints do not address the actual form of linear dependency present, and their use has resulted in many inconclusive and even misleading interpretations of age-period-cohort data.

3. THE MULTIPLE CLASSIFICATION MODEL: MATRIX FORMULATION

Our discussions to follow will be greatly facilitated by utilizing matrix algebra. To represent the multiple classification model (2.1) in matrix notation, we define the following matrices:

$\underset{\sim}{Y}' = (Y_{11}, Y_{12}, \ldots, Y_{1p}; \; Y_{21}, Y_{22}, \ldots, Y_{2p}; \ldots; Y_{a1}, Y_{a2}, \ldots, Y_{ap})$

is the vector of observed responses;

$\underset{\sim}{1}$ is an $(ap \times 1)$ column vector of ones;

$\underset{\sim}{A} = (\underset{\sim}{A}_1, \underset{\sim}{A}_2, \ldots, \underset{\sim}{A}_a)$ with $\underset{\sim}{A}_i$ an $(ap \times 1)$ column vector containing ones in positions corresponding to $Y_{i1}, Y_{i2}, \ldots, Y_{ip}$ and zeros elsewhere;

$\underset{\sim}{B} = (\underset{\sim}{B}_1, \underset{\sim}{B}_2, \ldots, \underset{\sim}{B}_p)$ with $\underset{\sim}{B}_j$ an $(ap \times 1)$ column vector containing ones in positions corresponding to $Y_{1j}, Y_{2j}, \ldots, Y_{aj}$ and zeros elsewhere;

$\underset{\sim}{C} = (\underset{\sim}{C}_1, \underset{\sim}{C}_2, \ldots, \underset{\sim}{C}_{a+p-1})$ with $\underset{\sim}{C}_k$ an $(ap \times 1)$ column vector containing ones in positions corresponding to all Y_{ij}'s for which $k = (a - i + j)$;

and

$$\underset{\sim}{\alpha}' = (\alpha_1, \alpha_2, \ldots, \alpha_a), \quad \underset{\sim}{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p), \quad \underset{\sim}{\gamma}' = (\gamma_1, \gamma_2, \ldots, \gamma_{a+p-1}).$$

Then, with $\underset{\sim}{X} = (\underset{\sim}{1}, \underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C})$ and $\underset{\sim}{\xi}' = (\mu, \underset{\sim}{\alpha}', \underset{\sim}{\beta}', \underset{\sim}{\gamma}')$, the multiple classification model (2.1), in matrix notation, has the form

$$\begin{aligned} E(\underset{\sim}{Y}) &= \underset{\sim}{1}\mu + \sum_{i=1}^{a} \underset{\sim}{A}_i \alpha_i + \sum_{j=1}^{p} \underset{\sim}{B}_j \beta_j + \sum_{k=1}^{a+p-1} \underset{\sim}{C}_k \gamma_k \\ &= \underset{\sim}{1}\mu + \underset{\sim\sim}{A}\underset{\sim}{\alpha} + \underset{\sim\sim}{B}\underset{\sim}{\beta} + \underset{\sim\sim}{C}\underset{\sim}{\gamma} = \underset{\sim\sim}{X}\underset{\sim}{\xi}. \end{aligned} \tag{3.1}$$

It has been shown by several researchers (see e.g., Portman [1963]) that the matrix $\underset{\sim}{X}$ is *four* less than full rank. Three of these four linear dependencies among the columns of $\underset{\sim}{X}$ are easily identified, namely,

$$\sum_{i=1}^{a} \underset{\sim}{A}_i - \underset{\sim}{1} = \sum_{j=1}^{p} \underset{\sim}{B}_j - \underset{\sim}{1} = \sum_{k=1}^{a+p-1} \underset{\sim}{C}_k - \underset{\sim}{1} = \underset{\sim}{0} . \tag{3.2}$$

The exact form of the fourth linear dependency has, up until now, not been identified. The following theorem rectifies that situation.

THEOREM: For model (3.1), the following linear relationship holds among the columns of $\underset{\sim}{X}$:

$$\sum_{i=1}^{a} \left[i - \frac{(a+1)}{2} \right] \underset{\sim}{A}_i - \sum_{j=1}^{p} \left[j - \frac{(p+1)}{2} \right] \underset{\sim}{B}_j + \sum_{k=1}^{a+p-1} \left[k - \frac{(a+p)}{2} \right] \underset{\sim}{C}_k = \underset{\sim}{0} . \tag{3.3}$$

Proof: Focus on the entry $E(Y_{ij})$ in $E(\underset{\sim}{Y})$. For this particular element, the relationship among the corresponding elements in (3.3) is

$$(1)\left[i - \frac{(a + 1)}{2}\right] - (1)\left[j - \frac{(p + 1)}{2}\right] + (1)\left[(a - i + j) - \frac{(a + p)}{2}\right] = 0,$$

and this completes the proof.

Some comments are in order regarding the structure of the constraint (3.3). Since the orthogonal polynomial values for assessing the *linear* effect of an equally-spaced variate with ℓ levels are given by the values of $\left\{i - \frac{(\ell + 1)}{2}\right\}$ for $i = 1,2,\ldots,\ell$, equation (3.3) says, loosely speaking, that "(the linear component of the age columns) - (the linear component of the period columns) + (the linear component of the cohort columns) equals zero." This result is intuitively very appealing when one relates it to the analogous constraint, A - P + C = 0, in the continuous variable case mentioned earlier.

Once having reparametrized (3.1) to eliminate the constraints (3.2), the equating of two effects to obtain a full rank design matrix clearly does *not* remove the adverse effect of the inherent linear dependency (3.3). On the contrary, it can be shown that the resulting full-rank matrix is very nearly singular (i.e., is highly ill-conditioned), indicating the presence of extreme multicollinearity. As documented by many researchers (see e.g., Mason et al. [1975]), such severe multicollinearity leads to highly unstable parameter estimates, and explains why previous workers in this area have obtained bizarre numerical results that vary wildly as a function of which effects are

assumed equal. In the next section, we will discuss a method of analysis which deals directly with the linear dependency (3.3), and which provides stable and accurate estimates of the age, period, and cohort effect parameters of interest.

4. LEAST SQUARES AND PRINCIPAL COMPONENT REGRESSION ANALYSIS OF MODEL (3.1).

Let us begin this section by reparametrizing model (3.1) into a more workable form. First, define $\bar{\alpha} = \frac{1}{a}\sum_{i=1}^{a}\alpha_i$, $\bar{\beta} = \frac{1}{p}\sum_{j=1}^{p}\beta_j$, and $\bar{\gamma} = \frac{1}{(a+p-1)}\sum_{k=1}^{a+p-1}\gamma_k$. With $\alpha_i^* = (\alpha_i - \bar{\alpha})$, $\beta_j^* = (\beta_j - \bar{\beta})$, $\gamma_k^* = (\gamma_k - \bar{\gamma})$, and $\mu^* = (\mu + \bar{\alpha} + \bar{\beta} + \bar{\gamma})$, then, using (3.2) and the equality $\sum_{i=1}^{a}\alpha_i^* = \sum_{j=1}^{p}\beta_j^* = \sum_{k=1}^{a+p-1}\gamma_k^* = 0$, it is easy to show that (3.1) can be written as

$$E(\underset{\sim}{Y}) = \underset{\sim}{1}\mu^* + \sum_{i=1}^{a-1}\underset{\sim}{A}_i^*\alpha_i^* + \sum_{j=1}^{p-1}\underset{\sim}{B}_j^*\beta_j^* + \sum_{k=1}^{a+p-2}\underset{\sim}{C}_k^*\gamma_k^*, \qquad (4.1)$$

where

$$\underset{\sim}{A}_i^* = (\underset{\sim}{A}_i - \underset{\sim}{A}_a),\ \underset{\sim}{B}_j^* = (\underset{\sim}{B}_j - \underset{\sim}{B}_p),\ \text{and}\ \underset{\sim}{C}_k^* = (\underset{\sim}{C}_k - \underset{\sim}{C}_{a+p-1}). \qquad (4.2)$$

Note that (4.1) has the usual structure of a model reparametrized to eliminate the standard analysis of variance constraints (3.2).

Model (4.1) is, of course, still one less than full rank. And, it is easy to show using (3.3) and (4.2) that the linear dependency existing among the columns of the design matrix for model (4.1) has the specific structure

$$\sum_{i=1}^{a-1}\left[i - \frac{(a+1)}{2}\right]\underset{\sim}{A}_i^* - \sum_{j=1}^{p-1}\left[j - \frac{(p+1)}{2}\right]\underset{\sim}{B}_j^* + \sum_{k=1}^{a+p-2}\left[k - \frac{(a+p)}{2}\right]\underset{\sim}{C}_k^* = \underset{\sim}{0}. \tag{4.3}$$

As mentioned earlier, in an attempt to estimate the parameters in (4.1), previous authors have equated two effects (e.g., have set $\gamma_1^* = \gamma_2^*$, say) to produce a full-rank model. The resulting set of estimates represents one of an infinite number of possible solutions to the normal equations based on (4.1), with each such solution corresponding to a particular choice of generalized inverse for the design matrix for (4.1). Such a generalized inverse solution will provide unbiased estimators only if the particular parametric constraint employed to solve the normal equations actually holds among the population parameter values. Since such a relationship will never be known to hold with certainty, preoccupation with the criterion of unbiasedness is of little practical value.

There is an even more compelling reason why equating two or more effects is not to be recommended. Even if the corresponding population values were approximately equal, such a generalized inverse solution would still produce unreliable estimates because of the presence of extreme multicollinearity due to (4.3). Such multicollinearity manifests itself in terms of an ill-conditioned design matrix with an eigenvalue very close to zero in value, leading to unstable estimates which are often quite far away from the true parameter values.

The reduction of multicollinearity has served as the motivation for the recent development of such biased estimation techniques as ridge and principal component regression (see Marquardt, 1970, for

an excellent review). The philosophy behind such methods is that it is worthwhile to introduce a slight bias into an estimation procedure if there is an accompanying large decrease in variance, thus leading to a significant reduction in mean square error relative to a *supposedly* unbiased estimation technique. Such a philosophy becomes all the more appealing when one accepts the fact that the form of the true underlying population model is never actually known, and that the validity of the claim of unbiasedness rests on such unavailable knowledge.

It is in this spirit that we propose an estimation technique which produces both stable and accurate estimates of the parameters in model (4.1). The application of our suggested estimation procedure to the age-period-cohort analysis of U.S. and British lung cancer mortality rates from 1931 to 1975 has led to remarkably reliable estimates, as documented via small-scale computer studies involving known population models. These results will be reported in separate communications in the epidemiologic literature.

The basis for our procedure involves reparametrizing model (4.1) so as to deal *directly* with the constraint (4.3). This will give yet another generalized inverse solution to the normal equations based on (4.1); however, in this instance, the resulting solution is *unique*. It can be shown that this uniquely specified generalized inverse solution is identical to the solution obtained by employing principal component regression analysis, a biased estimation procedure which deals with our multicollinearity problem by eliminating the principal component (or eigenvector) corresponding to the linear dependency (4.3). The principal component solution is, in fact, the Moore-Penrose generalized inverse solution (see Hocking, 1976, p. 33). The demonstration

of the equivalency of these two estimation procedures when an *exact* singularity (i.e., a zero eigenvalue) is present will be omitted for reasons of space, although the proof is available from the authors upon request. A somewhat related discussion of these concepts, which does not mention principal component regression analysis, has been given by Mazumdar et al. (1980).

To reparametrize (4.1) using (4.3), we employ the parametric constraint analogous to (4.3), namely,

$$\sum_{i=1}^{a-1}\left[i - \frac{(a+1)}{2}\right]\alpha_i^* - \sum_{j=1}^{p-1}\left[j - \frac{(p+1)}{2}\right]\beta_j^* + \sum_{k=1}^{a+p-2}\left[k - \frac{(a+p)}{2}\right]\gamma_k^* = 0. \tag{4.4}$$

The reparametrization is accomplished by expressing any one of the parameters (say, γ_1^*) in (4.4) as a function of the others, then substituting that expression for γ_1^* into (4.1), and finally rearranging to obtain the following reparametrized model:

$$\begin{aligned} E(\underset{\sim}{Y}) &= \underset{\sim}{1}\mu^* + \sum_{i=1}^{a-1}\left\{\underset{\sim}{A}_i^* + \frac{\left[i - \frac{(a+1)}{2}\right]}{\left[\frac{(a+p)}{2} - 1\right]}\underset{\sim}{C}_1^*\right\}\alpha_i^* \\ &+ \sum_{j=1}^{p-1}\left\{\underset{\sim}{B}_j^* - \frac{\left[j - \frac{(p+1)}{2}\right]}{\left[\frac{(a+p)}{2} - 1\right]}\underset{\sim}{C}_1^*\right\}\beta_j^* + \sum_{k=2}^{a+p-2}\left\{\underset{\sim}{C}_k^* + \frac{\left[k - \frac{(a+p)}{2}\right]}{\left[\frac{(a+p)}{2} - 1\right]}\underset{\sim}{C}_1^*\right\}\gamma_k^* \\ &= \underset{\sim}{X}^*\underset{\sim}{\xi}^*, \text{ say, where} \end{aligned} \tag{4.5}$$

$$\underset{\sim}{\xi}^{*\prime} = (\mu^*;\alpha_1^*,\ldots,\alpha_{a-1}^*;\beta_1^*,\ldots,\beta_{p-1}^*;\gamma_2^*,\ldots,\gamma_{a+p-2}^*).$$

Our estimator of $\underset{\sim}{\xi}^*$ is the standard least squares estimator

$$\hat{\underset{\sim}{\xi}}^* = (\underset{\sim}{X}^{*\prime}\underset{\sim}{X}^*)^{-1}\underset{\sim}{X}^{*\prime}\underset{\sim}{Y}, \tag{4.6}$$

with $\hat{\gamma}_1^*$ obtained from the constraint (4.4). The estimates so obtained are, of course, the same regardless of which parameter

is eliminated from (4.1) using (4.4).

As a simple example, the design matrix $\underset{\sim}{X}^*$ for the special case $a = 3$, $p = 4$ diagrammed in Section 2 is as follows based on (4.5):

$$\begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & \frac{7}{5} & 0 & -\frac{8}{5} & -\frac{6}{5} & -\frac{4}{5} & -\frac{2}{5} & -\frac{4}{5} & -\frac{6}{5} & -\frac{8}{5} \\
1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & -1 & -1 & -1 & 0 & 0 & 0 & 1 \\
1 & -\frac{7}{5} & -1 & \frac{8}{5} & \frac{1}{5} & -\frac{1}{5} & -\frac{3}{5} & -\frac{1}{5} & \frac{1}{5} & \frac{3}{5} \\
1 & -1 & -1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & -1 & -1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 1 & 0
\end{bmatrix}$$

The estimator $\hat{\underset{\sim}{\xi}}^*$ given by (4.6) is not an unbiased estimator of $\underset{\sim}{\xi}$ unless the constraint (4.4) actually holds among the population parameter values. However, limited numerical evaluations have suggested that the bias is small if a and p both exceed 5 in value. More detailed simulation studies are required regarding these bias considerations.

ACKNOWLEDGMENTS

The authors wish to thank Dean Bernard G. Greenberg for suggesting the problem and Dr. James E. Grizzle for some helpful suggestions during the course of the research. We also gratefully acknowledge grant support from the National Institute of Environmental Health Sciences and from The Council for Tobacco Research - U.S.A., Inc. (Special Project #102R1). A.M.D.G.

REFERENCES

Baltes, P. B. [1968]. Longitudinal and cross-sectional sequences in the study of age and generation effects. *Human Development* 11, 145-71.

Barrett, J. C. [1973]. Age, time and cohort factors in mortality from cancer of the cervix. *Journal of Hygiene, Camb.* 71, 253-9.

Greenberg, B. G., Wright, J. J. and Sheps, C. G. [1950]. A technique for analyzing some factors effecting the incidence of syphilis. *J. Amer. Statist. Assoc.* 45, 373-99.

Hocking, R. R. [1976]. The analysis and selection of variables in linear regression. *Biometrics* 32, 1-49.

Hussein, M. [1961]. A statistical study of factors influencing the incidence of breast cancer in females. Ph.D. Dissertation, University of North Carolina, Chapel Hill, North Carolina.

Knoke, D., and Hout, M. [1974]. Social and demographic factors in American political party affiliation, 1952-72. *American Sociological Review* 39, 700-13.

Marquardt, D. W. [1970]. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics* 12, 591-612.

Mason, K. O., Mason, W. M., Winsborough, H. H., and Poole, W. K. [1973]. Some methodological issues in cohort analysis of archival data. *American Sociological Review* 38, 242-58.

Mason, R. L., and Webster, J. T. [1975]. Regression analysis and problems of multicollinearity. *Comm. in Statist.* 4, 277-92.

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

Mazumdar, S., Li, C. C., and Bryce, G. R. [1980]. Correspondence between a linear restriction and a generalized inverse in linear model analysis. *The American Statistician* 34, 103-5.

Portman, R. M. [1963]. Estimation of time, age, and cohort effects. Ph. D. Dissertation. N. C. State University, Raleigh, North Carolina.

Schaie, K. W. [1965]. A general model for the study of development problems. *Psychological Bulletin* 64, 92-107.

Searle, S. R. [1971]. *Linear Models*. Wiley, New York.

Winsborough, H. H. [1975]. Age, period, cohort, and education effects on earnings by race - an experiment with a sequence of cross-sectional surveys. *Social Indicator Models*, Land, K. C., and Spillerman, S. (eds.) Russell Sage, New York.