IDENTIFICATION AND ESTIMATION OF AGE, PERIOD AND COHORT EFFECTS IN THE ANALYSIS OF DISCRETE ARCHIVAL DATA

Stephen E. Fienberg, University of Minnesota
William M. Mason, University of Michigan

1. INTRODUCTION

The discussion below summarizes the extended version of the paper, which exceeds the page limit for inclusion in the Proceedings.

This paper shows how to identify and estimate age, period and cohort effects in models which are log-additive with respect to a categorical response variable, and discusses the source of the identification problem in such models. Mason et al. [1973] have a similar intent, but focus on models with quantitative response variables.

In the general case we treat there are replicated cross-sectional data for each of J regularly spaced points in time. Our formal considerations apply to a wide variety of data. For convenience of reference, however, we shall suppose that our data are from sample surveys of individuals. Then, for each of the J points in time (hereafter periods) the data can be combined into age groups of respondents, e.g., 20-24 years of age. We assume that the range in time (e.g., years) covered by each age group equals the interval in time between successive periods for which we have data. Thus, all those in the i-th of I age groups in the j-th of J periods correspond to the same birth cohort as those in age group i+1 at the subsequent period, j+1. With I age groups and J periods there are K = I+J-1 cohorts.

We are interested in the effects of age, period, and cohort on a categorical response variable. For simplicity we begin with a dichotomous response variable. Let P_{ijk} denote the probability of a positive response given age i, period j and cohort k = i-j+J. Then one possible model of interest is the logistic response model:

$$\Omega_{ijk} = \log \frac{P_{ijk}}{1 - P_{ijk}}$$

$$= W + W_{1(i)} + W_{2(j)} + W_{3(i-j+J)} \qquad (1)$$

where the subscripted parameters in (1) are deviations from W, i.e.,

$$\Sigma_i W_{1(i)} = \Sigma_j W_{2(j)} = \Sigma_k W_{3(k)} = 0. \qquad (2)$$

This model postulates simultaneous age, period and cohort effects on the log-odds (or logit) of the probability of success. The notation we use here is consistent with that in Bishop, Fienberg and Holland [1975]. The model can be expanded to include further explanatory variables (e.g., sex, race, socioeconomic status) as well as their interaction effects with age, period and cohort. The model is directly analogous to the age-period-cohort model for quantitative response variables discussed by Mason et al. [1973].

The problems of identification and estimation of model (1)-(2), and ways of arraying the data pertinent to model (1)-(2), are most simply illustrated for the case in which there are data from 3 periods, 3 age groups and, therefore, 5 cohorts. The extended version of the paper subjects the 3x3x2 case to detailed analysis, and then proceeds to the IxJx2 general case. Here, we shall illustrate one way of arraying the data for the case of 3 age groups and 3 periods, and then state the results for the general case in compressed form.

For the 3-age group and 3-period case, the basic data with which to formulate, estimate and assess the adequacy of an age-period-cohort model come from a series of 3 surveys, one for each period. The data then consist of counts $\{x_{ijk\ell}\}$, where $\ell = 1$ corresponds to a positive response and $\ell = 2$ to a negative response, and thus the counts form a 3x3x2 cross-classification with the sample sizes (marginal configurations $\{x_{+j++}\}$) fixed by design, as depicted in Table 1. It is important to bear in mind that one of the first three subscripts is redundant since k = i-j+3.

If we define the expected cell values corresponding to Table 1 under the logistic response model (1)-(2) by $\{m_{ijk\ell}\}$ we have:

$$m_{ijk1} = x_{ijk+} P_{ijk} \qquad (3)$$

and

$$m_{ijk2} = x_{ijk+}(1 - P_{ijk}) \ . \qquad (4)$$

The basic logistic response model can now be written in terms of expected cell values as:

$$\Omega_{ijk} = \log \frac{m_{ijk1}}{m_{ijk2}}$$

$$= W + W_{1(i)} + W_{2(j)} + W_{3(i-j+J)} \ , \qquad (5)$$

and analyses involving this model treat the marginal configuration $\{x_{ijk+}\}$ as fixed, even though only the totals $\{x_{+j++}\}$ are fixed by design.

Table 1 is just one of several ways in which the data can be arrayed. Under certain circumstances it may be preferable to construct age by cohort or period by cohort tables. Such alternative tables are also used in the extended version of the paper to aid the exposition of the identification problem.

2. IxJx2 CASE

A. Identification

Because cohort is determined uniquely by age

and period, we must take care to ensure that the unique effects of age, period and cohort in the logistic response model are estimable. In the linear model for quantitative response variables analogous to (1)-(2) it is well known (Mason et al. [1973]) that all of the parameters are not estimable. The same is the case for the logistic response model.

In the case of I age groups and J periods it might be supposed that there are I-1 independent parameters for the effect of age on the response variable, J-1 for the period effects, and I+J-2 for the cohort effects for a total of 2I+2J-4. It turns out, however, that the number of independent effect parameters is 2I+2J-5, which is one less than specified by the basic logistic response model, (1)-(2). Thus, not all of the effect parameters specified by model (1)-(2) are identified.

The source of the identification problem can be described in various ways. One useful insight is that the effects of age, period and cohort contain linear and higher order components (e.g., quadratic components in the 3x3x2 case). It can be shown that the linear component of any one set of effect parameters (e.g., those of age) can not be separated from the linear components of the other two sets of parameters.

In order to estimate all of the independent effect parameters of model (1)-(2) it is necessary to put a single restriction on the model, e.g., $W_{1(1)} = W_{1(3)}$, $W_{2(1)}$ = constant, or $W_{3(1)} = W_{3(2)}$. We refer to such a restriction as an identification specification. Different identification specifications lead to effectively different models. An identification specification is like any other assumption in a statistical model that is not capable of direct verification as part of an analysis; it must be grounded in substantive theory relating to the data in question or it must come from observations on and analyses of other data on related phenomena.

Although the technical aim of making an identification specification is to allow the estimation of the linear components of the effect parameters, the most reasonable types of specifications are likely to be that two or more age groups, periods or cohorts have the same effects on the log-odds-ratios. What is more interesting from a substantive point of view than specifying a single identification specification, is the specification of overidentifying restrictions on the effect parameters based on considerable collateral information. When the resulting model fits the data well, not only do we solve the identification problem but we also get some verification of hypotheses related to the substantive theory.

B. Estimation

Given the logistic response model (1)-(2) and J independent simple random samples at 3 properly spaced points in time, and given an identification specification, it is possible to obtain maximum likelihood estimates of the effect parameters and of the expected values $\{m_{ijk\ell}\}$ corresponding to the observed frequencies $\{x_{ijk\ell}\}$. The likelihood equations can be solved by the Newton-Raphson iterative procedure (Bock [1975], Haberman [1974]). Alternatively, for sufficiently simple identification specifications the likelihood equations can be solved using iterative proportional fitting (Bishop, Fienberg and Holland [1975]), or for more complex identification specifications using generalized iterative proportional fitting (Darroch and Ratcliff [1972]). If (generalized) iterative proportional fitting is used, one solves first for the $\{\hat{m}_{ijk\ell}\}$ and then for the estimated effect parameters. If the Newton-Raphson method is used, one solves first for the estimated effect parameters. Thus, use of the Newton-Raphson method requires prior resolution of the identification problem. This method has been programmed for general purpose work with discrete data (Bock and Yates, [1973]) and is preferable to iterative scaling for several reasons. The extended version of the paper discusses estimation in more detail than is possible here.

C. Degrees of Freedom

Degrees of freedom equal the number of conditional log-odds for the response variable minus the number of independent effect parameters minus one (for the grand mean). For the "full" model, i.e., the model in which only one identification specification has been made, there are (I-2)(J-2) degrees of freedom. Table 2 lists the full model and the 7 possible reduced models (to be discussed below), and the associated minimal sufficient statistics and degrees of freedom. Goodman [1975] gives a similar table.

D. Goodness of Fit

Once we have estimated expected cell values, we can test the goodness-of-fit of model (1)-(2) using either the Pearson statistic,

$$X^2 = \Sigma \frac{(x_{ijk\ell} - \hat{m}_{ijk\ell})^2}{\hat{m}_{ijk\ell}} , \qquad (6)$$

or the likelihood ratio statistic,

$$G^2 = 2\Sigma x_{ijk\ell} \log \frac{x_{ijk\ell}}{\hat{m}_{ijk\ell}} . \qquad (7)$$

If the model is correct then either statistic is asymptotically distributed as a chi-square variate with degrees of freedom determined as described above.

E. Reduced Models

If the logistic response model with age, period and cohort effects, and with an associated identification specification, provides an acceptable fit to the data, then we would logically wish to explore whether only two sets of effects may suffice, i.e., whether we can equate one set of effects (age or period or cohort) to zero. Fitting such reduced models is a straightforward

task with any computer program designed to fit standard loglinear models to multidimensional arrays (with or without structural zeros). There is no longer an identification problem when we deal with reduced models because there is no way for a linear component for one type of effect to become confounded with the other two types.

The fit of reduced models can be assessed using the standard goodness-of-fit statistics, (6) and (7), and we can compare the fit of the reduced models to the specified age-period-cohort models using the log-likelihood-ratio statistic for nested models, i.e., the conditional likelihood ratio test for the fit of the reduced model given that the age-period-cohort model is correct. In a similar fashion we can fit reduced models with only one set of effect parameters. Degrees of freedom and other information for reduced models are given in Table 2.

3. MULTIVARIATE QUALITATIVE RESPONSE MODELS

We have thus far dealt with logistic response models measuring the effects of age, period, and cohort on a dichotomous response variate. Here we consider two extensions of these models for polytomous response variates. Suppose the response variable has L categories, and the basic data array is the IxJxL age-period-response table (alternatively, the Ix(I+J-1)xL age-cohort-response table, or the Jx(I+J-1)xL period-cohort-response table). As before, we label the counts in the array using four subscripts, i.e., $\{x_{ijk\ell}\}$ where $\ell = 1,2,\ldots, L$, and we denote the corresponding expected cell values by $\{m_{ijk\ell}\}$. When we had a dichotomous response variate we considered a model for the single logit structure, $\log(m_{ijk1}/m_{ijk2})$. Now that we have L categories for our response variable we would like to consider models for L-1 different logit structures. Two possible ways to define these are:

$$\text{(i)}\quad \log\left(\frac{m_{ijk\ell}}{\sum_{\ell'>\ell} m_{ijk\ell'}}\right) \quad \text{for } \ell = 1,2,\ldots, L-1\ , \qquad (8)$$

and

$$\text{(ii)}\quad \log\left(\frac{m_{ijk\ell}}{m_{ijk\ell+1}}\right) \quad \text{for } \ell = 1,2,\ldots, L-1\ . \qquad (9)$$

If we would like to fit models with the same parametric structure to the L-1 logits, and to have them correspond as a group to a loglinear model for the $\{m_{ijk\ell}\}$, then our choice would be (9). (Note that (9), in such circumstances, is equivalent to models with the same parametric structure for the L-1 logits

$$\text{(iii)}\quad \log\left(\frac{m_{ijk\ell}}{m_{ijkL}}\right) \quad \text{for } \ell = 1,2,\ldots, L-1\ , \qquad (10)$$

or any other set of L-1 logits involving the logarithm of the odds for pairs of expected values). The generalization of the logistic response model we would consider for (9) is:

$$\Omega_{ijk\ell} = \log\left(\frac{m_{ijk\ell}}{m_{ijk\ell+1}}\right)$$

$$= W^{(\ell)} + W^{(\ell)}_{1(i)} + W^{(\ell)}_{2(j)} + W^{(\ell)}_{3(i-j+J)}\ , \qquad (11)$$

(for $\ell = 1,2,\ldots, L-1$),

where

$$\sum_i W^{(\ell)}_{1(i)} = \sum_j W^{(\ell)}_{2(j)} = \sum_k W^{(\ell)}_{3(k)} = 0\ . \qquad (12)$$

All of the earlier results for the dichotomous response variate carry over immediately to this set of models as long as we make sure that (a) summations involving the fourth subscript run up to L instead of 2, and (b) the degrees of freedom listed in Table 2 are all multiplied by L-1.

If the L response categories are ordered and it makes substantive sense to think of the effects linking the response variable to age, period, and cohort as increasing linearly with the category number (e.g., the category number represents some latent variable), then we may wish to test for the equality of various effect parameters, e.g.,

$$W^{(\ell)}_{1(i)} = W^{*}_{1(i)} \quad \text{for } \ell = 1,2,\ldots, L-1\ . \qquad (13)$$

Such reduced models can be handled without trouble using the methodology of loglinear models with ordered categories for some of the variables (see, e.g., Fienberg [1977]).

When the response categories have a natural order, e.g., educational attainment (grade school, high school, college, graduate school), the other choice of logits, in expression (8) may be preferable. The quantities $(m_{ijk\ell}/\sum_{\ell'>\ell} m_{ijk\ell'})$ are often referred to as continuation ratios, and they are of substantive interest in various fields. There is also a technical reason for working with the logits in expression (8). Let $P_{ijk\ell}$ be the probability of a response in category ℓ given age i and period j, where $\Sigma_\ell P_{ijk\ell} = 1$. Then, when the $\{x_{ijk\ell}\}$ consist of observations from IJ independent multinomial variables with sample sizes $\{x_{ijk+}\}$ and cell probabilities $\{P_{ijk\ell}\}$,

$$m_{ijk\ell} = x_{ijk+}\, P_{ijk\ell}\ , \qquad (14)$$

so that

$$\frac{m_{ijk\ell}}{\sum_{\ell'>\ell} m_{ijk\ell'}} = \frac{P_{ijk\ell}}{\sum_{\ell'>\ell} P_{ijk\ell'}}\ . \qquad (15)$$

We can write the multinominal likelihood functions as products of L-1 binomial likelihoods, the ℓ-th of which has sample sizes $\{\sum_{\ell' \geq \ell} x_{ijk\ell'}\}$ and cell probabilities $\{P_{ijk\ell} / \sum_{\ell' \geq \ell} P_{ijk\ell'}\}$. This means that if we use the method of maximum likelihood to estimate the parameters in the logistic response models

$$\log\left(\frac{m_{ijk\ell}}{\sum_{\ell'>\ell} m_{ijk\ell}}\right) = W^{(\ell)} + W^{(\ell)}_{1(i)} + W^{(\ell)}_{2(j)} + W^{(\ell)}_{3(i-j+J)} , \qquad (16)$$

(for $\ell = 1,2,\ldots, L-1$) ,

subject to (12), then we can do the estimation separately for each logit model using methods applicable to dichotomous response variates, and we can simply add individual chi-square statistics to get an overall goodness-of-fit statistic for the set of models. Moreover, the observed binomial proportions

$$x_{ijk\ell} / \sum_{\ell' \geq \ell} x_{ijk\ell'} , \qquad \ell = 1,2,\ldots, L-1 , \qquad (17)$$

are asymptotically independent of each other so that we can assess the fit to the L-1 logit models, and various associated reduced models, independently.

For the logistic response models in (16) it might be of substantive interest to explore the equality of parameters across models as in expression (13). The estimated expected values for such models and the associated tests of fit can be handled with each by thinking in terms of a set of counts with 5 subscripts, $\{y_{ijk\ell t}\}$ where

$$y_{ijk\ell t} = \begin{cases} x_{ijk\ell} & \text{for } t = 1 \\ \sum_{\ell'>\ell} x_{ijk\ell'} & \text{for } t = 2 \ . \end{cases} \qquad (18)$$

Now, if we let $m^{*}_{ijk\ell t}$ be the expected value under model (16) corresponding to $y_{ijk\ell t}$, then we fit the L-1 models simultaneously by fitting a hierarchical loglinear model to the $\{y_{ijk\ell t}\}$ with minimal sufficient statistics $\{y_{ijk+t}\}$, $\{y_{i++\ell t}\}$, $\{y_{+j+\ell t}\}$, $\{y_{++k\ell t}\}$. If we restrict the model so that (13) holds we fit the loglinear model with minimal sufficient statistics $\{y_{ijk+t}\}$, $\{y_{i++\ell+}\}$, $\{y_{+j+\ell t}\}$, $\{y_{++k\ell t}\}$. We can also handle similar reduced models involving equality of period and cohort effects across the L-1 logistic response structures.

4. OTHER EXTENSIONS

In the detailed paper we consider the identification problem for a subset of the cases in which the age group intervals are not identical to the period intervals. We examine in particular the case of r-year (e.g., 5-year) age groups and 2r-year (e.g., 10-year) periods. Once an identification specification has been made for this case, estimation can proceed as described earlier.

5. EXAMPLE

The extended version of the paper includes a detailed analysis of the educational attainment of white males. The data are from the U.S. Decennial Censuses of 1940-1970. Education is treated as a set of continuation ratios, and an overidentified age-period-cohort model is fit to the various logged continuation ratios. Reduced models are also fit and discussed.

ACKNOWLEDGEMENT

We are grateful to Michael Battaglia for able and devoted computational assistance.

REFERENCES

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. [1975]. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Massachusetts: The MIT Press.

Bock, R. D. [1975]. Multivariate Statistical Methods in Behavioral Research. New York: McGraw-Hill.

Bock, R. D., and Yates, G. [1973]. MULTIQUAL: Log-linear Analysis of Nominal or Ordinal Qualitative Data by the Method of Maximum Likelihood--A Fortran IV Program. Chicago: National Educational Resources.

Darroch, J. N., and Ratcliff, D. [1972]. "Generalized iterative scaling for log-linear models," The Annals of Mathematical Statistics, 43, 1470-80.

Fienberg, S. E. [1977]. The Analysis of Cross-Classified Data. Cambridge, Massachusetts: The MIT Press.

Goodman, L. A. [1975]. "A note on cohort analysis using multiplicative models and the modified multiple regression approach." Unpublished manuscript.

Haberman, S. J. [1974]. The Analysis of Frequency Data. Chicago: University of Chicago Press.

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.

Table 1: Age by Period Display

Positive Response

Age \ Period	1	2	3
1	x_{1131}	x_{1221}	x_{1311}
2	x_{2141}	x_{2231}	x_{2321}
3	x_{3151}	x_{3241}	x_{3331}

Negative Response

Age \ Period	1	2	3
1	x_{1132}	x_{1222}	x_{1312}
2	x_{2142}	x_{2232}	x_{2322}
3	x_{3152}	x_{3242}	x_{3332}

Table 2: Information Associated with Model (1), and Various Reduced Models

Subscripted Logistic Parameters in Model (1)	Degrees of Freedom	Minimal Sufficient Statistics*
1. None	IJ-1	$\{x_{+++\ell}\}$
2. Age	I(J-1)	$\{x_{i++\ell}\}$
3. Period	J(I-1)	$\{x_{+j+\ell}\}$
4. Cohort	(I-1)(J-1)	$\{x_{++k\ell}\}$
5. Age, Period	(I-1)(J-1)	$\{x_{i++\ell}\}$, $\{x_{+j+\ell}\}$
6. Age, Cohort	(I-1)(J-2)	$\{x_{i++\ell}\}$, $\{x_{++k\ell}\}$
7. Period, Cohort	(I-2)(J-1)	$\{x_{+j+\ell}\}$, $\{x_{++k\ell}\}$
8. Age, Period, Cohort	(I-2)(J-2)	$\{x_{i++\ell}\}$, $\{x_{+j+\ell}\}$, $\{x_{++k\ell}\}$

*For each model we always include the totals, $\{x_{ijk+}\}$, implied by the logistic structure, as well as the statistics listed.