Some Age-Period-Cohort models to study Time Trends In the Incidence of Cancer of the Larynx

Khudadad Khan
Institute of statistics
University of the Punjab
Quaid-e-Azarn Campus, Lahore

ABSTRACT

In this paper, three approaches are described to fit Age-Period-Cohort models to study the temporal variation in the incidence of cancer Osmond and Gardner (1982) estimate the parameters for the full models from the estimates of three two-factor models minimizing the Euclidean distance between the parameters. The concept of drift given by Clayton and Schifflers (1987) is described. Robertson and Boyle explained how individual records could be used to get estimates of parameters for the full Age-Period-Cohort model with more precision. The methods of Osmond and Gardner (1982) and Decarli and La Vecchia (1987) are conceptually similar. The parametric bootstrap method can be applied to provide Monte Carlo estimates of the standard errors.

1. INTRODUCTION

The data on incidences of cancer of the larynx can be used to study the temporal variation of incidence by analyzing the set of rates arranged in two–way tables of age at incidence by time and periods of incidence. Birth cohort is another factor, which can play an important role in producing influencing the treads. Age, period and cohort models are the usual approach for estimating temporal variation. Here, three approaches to analyze the data by age, period and cohort models are reviewed. These approaches are given by Robertson and Boyle (1986), Clayton and Schifflers (1987) and Decarli and La Vecchia (1987) Poisson regression can be used to meet this purpose. The technique given by Decarli and La Vecchia does not provided information to calculate the standard errors of the estimates directly. A parametric bootstrap method can be used to calculate the standard errors of the estimates.

2 MATERIALS AND METHODS

A number of different methods are available to estimate the separate effects of age, period and cohort. Some require the use of a three–way table while others require a two–way table. The temporal variation through age groups and time periods can be studied without making any condition of grouping or any approximation as the age group and time periods can be exclusively defined. To study the temporal variation through cohorts it must be assumed that the effect of overlapping is negligible.

Usually published data are available in 2–way tables. With individual records it is possible to construct a 3–way table of age group by period by birth cohort. For 3–way tables, each age, period cell has two cohorts associated with it. This table will increase the number of cohorts by one. In this case overlapping still exists but this time the cohorts overlap for only one year. The effect of one year overlapping can quite safely be ignored (Robertson and Boyle, 1986).

New approaches to the analysis of temporal variation in disease incidence lead to the generalization of indirect standardization to the estimation of parameters of the age–period–cohort model (Holford, 1983). The parameters of the age–period–cohort model can also be estimated through Poisson or logistic regression analysis. In the following these techniques are discussed. Additionally the implementation of age–period–cohort model approach in the well–known statistical package GLIM is considered (Decarli and La Vecchia, 1987).

The GLIM macro provided by Decarli and La Vecchia does not calculate standard errors of the estimates. There is no information in their paper about the calculation of the standard errors. The estimates of the age–period–cohort model by the method proposed by Decarli and La Vecchi are based on the minimization of a penalty function. The derivative of the penalty function is not linear. So it is not easy to calculate the standard errors analytically. Hence a parametric bootstrap method can be used to provide Monte Carlo estimates.

3 MODELS

The following notation is used.

m = number of 5-year age groups (=10)

n = number of 5-year periods (=6)

YVAR = n_{ij}

ERROR = Poisson

LINK = log

If some of the values of the parameters for the linear prediction model are fixed in advance then an offset Is required (Glim Manual, 1985). Here the offset will be log (Y_{ij}) as

$$\log E(r_{ij}) = \log E(\frac{n_{ij}}{y_{ij}})$$

$$= \log E(n_{ij}) = \log(Y_{ij})$$

The major problem of age, period, and cohort models is that these three factors are not independent since availability of any 'two implies knowledge of the third. The three factors are combined with the relation

Birth cohort = Time Period – Age.

Thus exists a linear dependency among these three factors (Holford, 1983). However they may exert a simultaneous influence in that they index contributory causal factors (Osmond and Gardner, 1982)

Many authors have discussed the problem of identifiability. Fienberg and Mason (1978) have shown that a single additional constraint can ensure the uniqueness of the estimates provided some valid and reasonable prior information is available. The additional constraint could be like that the age group one has the same effect as age group two or the last time period has the same effect as its previous time period has or that the first cohort has same effect as the last cohort, etc. But there is another problem with the addition of the constraint. Different constraints will yield different estimates. Hence this is not a recommendable approach which can be recommended.

3.1 Osmond and Gardner technique

Osmond and Gardner (1982) have discussed the problem of identification of the parameters for the full age, period and cohort model. Let us define three new parameters α_i*, β_j* and γ_k* as

$$\alpha i^* = \alpha i + \lambda(m-i)$$

$$\beta j^* = \beta j + \lambda j$$

$$\gamma k^* = \gamma k - \mu k$$

where λ is constants and $k = m - i + j$ by changing the values of λ we can get different estimates of αi^*, βj^* and γk^* but the fitted values will be ıne same all the time as

$$\alpha i^* = \alpha i + \lambda(m-i)$$

$$\beta j^* = \beta j + \lambda j$$

$$\lambda k^* = \lambda k - \lambda k$$

where λ is constants and $k = m - i + j$. By changing the values of λ we can get different estimates of αi&, βj^* and γk^* but the fitted values will be the same all the time as

$$\alpha i^* + \beta j^* + \lambda k^* = \alpha i + \beta j + \gamma_k.$$

As a consequence, no unique solution is available Osmond and Gardner (1982) have overcome the difficulty of the linear dependency by fitting three possible two–factors models

$$\log E\,(r_{ij}) = \mu + \alpha_i + \beta_j$$

$$\log E\,(r_{ij}) = \mu + \alpha_i + \gamma_k$$

$$\log E(r_{ij}) = \mu + \beta_j + \gamma_k.$$

where α_i represents the age effect and γk is the cohort effect and represents the differences between parallel curves for age-cohort model. The multiplicative model is defined as

$$r_{ik} = \alpha_i \ \gamma_k$$

where $\alpha_i`$ is the antilogarithms of $\alpha_j$ and $\lambda`_k$ is the antilogarithm of γk. Estimation of the cohort effects can be obtained in a similar fashion to that for the period effects. However the number of parameters increases. The cohort effects have an interpretation, which is similar to the period effects. Again it could be useful to focus attention on regular trend by reporting the first differences of the cohort effects — $(\gamma_2 - \gamma_1)_1$ $(\gamma_3 - \gamma_2)_1$ etc. The antilogarithms of these differences give the relative risks between adjacent cohorts.

If the age-period model fits the given data and there are reasons to believe that, for the age-period curve, the difference between the parallelism of the curves for time period justifies the assumption that all differences β_2–β_1, β_3–β_2, etc. are same, then it can be denoted by β_p say, i.e. there is equal difference between the parallel curves. For the age cohort model the same argument can be applied and it can be assumed that the differences $(\gamma_2 - \gamma_1)$, $(\gamma_3 - \gamma_2)$, etc. are equal and have a single value γ'c say.

Clayton and Schifflers have given an analysis of data on mortality from lung cancer in females in Belgium, during the period 1955 to 1978. Both the age-period model and age-cohort model seem to be good fits indicating a temporal variation which could equally be described by period or cohort. Clayton and Schifflers have named this variation as 'drift'. They defined the drift as a linear effect of period or cohort. They suggested log-linear models for age-period and age-cohort models involving drift parameter δ say.

The age drift model involving time periods can be written as

$$\log (r_{ij}) = \alpha_j + \delta(j{-}{-}n_0)$$

Here n_0 is the reference period, α_i are the fitted age-specific rates in the reference period and δ, the drift parameter is the constant change in log-rates from one period to the next. Clayton and Schifflers have suggested that α_i should be called 'cross-sectional' age.

If the drift is calculated from cohorts instead of periods the age drift model can be written as

$$\log\,(r_{lk}) = \alpha_i{}^* + \delta(k\text{-}c_0)$$

where c0 is the reference cohort $\alpha_i{}^*$ are the fitted age-specific rates in the reference cohort and δ (drift parameter) is the constant change in log-rates from one cohort to the next. Here, according to Clayton and Schiffjers, $\alpha_i{}^*$ would be called as the 'longitudinal' age. The two models are not the same. However the above two models give the same predictions for the rates. The models are indistinguishable until the relation between incidence and age is known and is not estimated.

According to Clayton and Schifflers, consideration of either specifically age-period or age-cohort models is justified only if the so-called drift model does not adequately describe the data.

Further if both age-period and age-cohort models do not fit the data adequately, both cohort and period effects should be included. Thus the age-period-cohort model, finally, includes

(i) Age parameters

(ii) Drift

(iii) Non-drift period effects

(iv) Non-drift cohort effects.

The full drift model can be written in one of the following forms

$$\log\,(r_{ij}) = \alpha_i + \delta_p + \beta j + \gamma k$$

$$\log\,(r_{ij}) = \alpha_i{}^* + \delta c + \beta j + \gamma_k$$

where P and C are the means of period groups and means of cohort groups. Other terms are as defined earlier. Here the first period is the reference period and the first cohort is the reference cohort.

From the non-drift period effects βj the amount of curvature for time periods period, 1960-1979. They established the independent effects of age, calendar time, and birth cohort on the observed pattern of incidence by adopting an alternative approach to this classical problem, which can be employed when data are available on individual records. However the assumption of a common age effect across the cohorts may not be completely valid (Clayton and Schifflers, 1987).

4. Summary and another approach.

The concept of temporal variation is introduced in section 1. Age, Period and Cohort model is defined in section 3 as

$$\log E(r_{ij}) = \log E\left[\frac{n_{ij}}{y_{ij}}\right] = \mu + \alpha_i + \beta_j + \gamma_k$$

where

$E(r_{ij})$ is the expected risk for a person in age group *i* in period *j*, αi, *1, ..., .m,* represents the age group effects, βj , j= *1,...,* *n,* the period effects and γk, k= 1, *:..,* c, the birth cohort effect, $\{n_{ij}\}$, *(i = 1, ..., m, j = 1, ..., n)* is the matrix of numbers of incidences or deaths in j^th^ age group and j^th^ period, $\{y_{ij}\}$, *(i = 1,..., m, j = 1, ..., n)* is the matrix of person-years at risk in j^th^ age group and j^th^ period and $\{r_{ij}\}$, *(i = 1,..., m, j = 1, ..., n) is the matrix of incidence rtes in i^th^ age group and j^th^ period*

Section 3.1 described the model presented by Osmond and Gardner (1982). Osmond and Gardner estimate the parameters for, the full models from the estimates of, three two factor models minimizing the Uclidean distance between the parameters. In section 3.2 the concept of drift given by Clayton and Schifflers, 1987 is described. Drift is a linear parameter, which gives the trends along Time Period or Birth Cohort. According the Clayton and Schifflers, the non-linear, factors should be involved only if the linear parameter *i.e.* drift is unable to represent the data. The macro given by Decarli and La Vecchia is discussed In section 3.4 Decarli and La Vecchia model is full Age, Period and Cohort model with error component. Robertson and Boyle (1987) explained how individual records could be used to get estimates of parameters for the full Age-Period-Cohort model with more precision (section 3.4) Three-way table can be constructed for data with individual record and one degree of freedom is increased. With addition of one degree of freedom there remains no need of putting additional restriction on the model.

The identifiability problem can only be overcome by using extra information or constraints. The methods of Osmond and Gardner (1982) and Decarli and La Vecchia (1987) are conceptually similar. The method of Decarli and La Vecchia relies on the minimization of a penalty function. The result is that we get no standard errors. Hence the precision of the estimated effects cannot be assessed.

The parametric bootstrap method can be applied to provide Monte Carlo estimates of the standard errors. The method of Robertson and Boyle (1986, 1987) uses extra information to form a 3-way table. One degree of freedom is increased by the addition of one cohort and there remains no need of additional constraint. The method of Clayton and Schifflers, (1987) is mathematically correct but the curvatures and drift are not easy to interpret. The method of Holford, (1983) is similar to that of Clayton and Schifflers. Another techniques to solve the identifiability problem for Age–Period-Cohort models is described below.

James and Segal (1982) discussed a model proposed by Moolgavkar et al., (1979). The model is

$$\log E\left(\frac{n_{ij}}{y_{ij}}\right) = \mu + \alpha_i + \gamma_k + \beta_j \delta_i$$

log where δ_i is the period interaction effect and reflects changes in the relative risk for two ages at different times. James and Segal have described the method for fitting this model. In this model the Interaction factor is included without including the main, factor β_j. McCullagh and Nelder (1983) caution against the use of models using interaction factors without main factors involved in the interaction factors.