CHAPTER 5

FITTING MODELS TO CONTINUOUS DATA

Grouping of cohort data into a multidimensional classification of cases/deaths and person-years by categories of age, calendar period, cumulative exposures indices and other fixed or time-varying factors is a convenient way of reducing a frequently massive set of information into a form suitable for statistical analysis. It encourages the investigator to examine disease rates calculated within each cell of the cross-classification, making plots of rates against quantitative exposure measurements for purposes of model development. Inferences regarding disease mechanisms are made possible by examining the data for trends in excess or relative risk measures according to ordered categories of age at onset of exposure, duration of exposure, or time since cessation of exposure. By assigning average duration or dose levels to these categories, quantitative regression models may be fitted for purposes of risk assessment. Validation of the fitted models is facilitated by the calculation of standardized differences (residuals) between observed and fitted numbers of cases in each cell.

We believe that such grouped data analyses are generally the method of choice for cohort analysis. Given the inherent limitations of cohort data in terms of the number of cases and the accuracy of recorded exposure variables, more elaborate approaches such as those embodied in the continuous data analyses that we now describe are perhaps best limited to special situations. A possible exception is the use of the method of case-control within a cohort sampling (§5.4) to conduct preliminary exploratory analyses in order to select variables for a final analysis, which is carried out using either grouped or continuous data techniques.

Restrictions on grouped data analyses

A key assumption of the grouped data approach is that disease rates are constant within each cell of the multidimensional cross-classification. While clearly an approximation, this assumption can be made more plausible by refining the classification, for example, by using five-year rather than ten-year intervals of age and calendar time. However, there are obvious restrictions on the number of different variables that can be considered simultaneously and on the number of levels or categories into which each variable is factored. When most cells contain few, if any, cases, the previously cited measures of goodness-of-fit based on comparisons of observed and expected (fitted) numbers of cases have little if any value. Practical difficulties arise in coping with large numbers of cells and estimating large numbers of parameters.

Scope of Chapter 5

This chapter develops methods of continuous cohort data analysis that utilize age, time and exposure measurements in their original form rather than after partitioning the data into discrete categories. Many different explanatory variables may be considered simultaneously in the same analysis. To a large extent, the methods presented here are applications and refinements of survival analysis techniques originally proposed by Cox (1972) and developed further in texts by Kalbfleisch and Prentice (1980) and Cox and Oakes (1984), which should be consulted for a more detailed development. We first review, in §5.1, the fundamental concept of a disease incidence rate, considered as a continuous function of age and/or time. We describe how model equations already developed to express the effects of exposure on disease rates calculated from grouped data are adapted to the continuous case. Section 5.2 introduces the 'partial likelihood' methodology for estimating regression coefficients in models in which the exposure variables are assumed to act multiplicatively on the background rates. It contains a detailed, worked example for the simplest situation – that of a single, binary (but age-dependent) exposure variable. In §5.3 we develop nonparametric estimates of unknown baseline disease rates, both for homogeneous samples and for heterogeneous ones in which the heterogeneity is expressed by covariables in the multiplicative model. When the background rates are determined from vital statistics or are assumed to have a specific parametric form, the same techniques provide a nonparametric description of how relative mortality rates (SMRs) may vary continuously with time since first exposure, time since cessation of exposure, or with some other relevant time variable. Plots of baseline or relative mortality functions against one or more time-varying factors are shown to be quite useful as a means of informally examining model assumptions. In §5.4, we present details about the 'case-control within a cohort' or 'synthetic retrospective study' sampling technique that was introduced in Chapter 1 as a device for conducting efficient, exploratory analyses of continuous cohort data. This section also presents analytical methods for gauging the influence of individual cases or controls on the estimated regression coefficients. In sections 5.5 and 5.6 these methods are applied systematically to the studies of Montana smelter workers and Welsh nickel refinery workers, and comparisons are made with results of grouped analyses of these same data sets already presented in Chapter 4.

5.1 Fundamentals of continuous data analysis

Continuous data methods rest fundamentally on the concept of an instantaneous disease rate considered as a continuous function of a continuous time variable t (see Chapter 2 of Volume 1). Let $\lambda(t)$ denote the rate for a given subject at time t such that $\lambda(t)\,dt$ is the probability of disease diagnosis or death in the time interval $(t, t+dt)$, given that he was alive and/or disease-free at its start. We assume there is a background rate function $\lambda_0(t)$ that represents the degree of risk for someone with no exposure or, in some cases, a standard set of exposures. The object of the data analysis is to construct models that describe how the exposure variables $\mathbf{x}(t)$, which may

themselves vary continuously and depend on time, act to modify the background rates $\lambda_0(t)$. Exposure effects are expressed parametrically in terms of a vector $\boldsymbol{\beta}$ of unknown parameters, and the statistical problem is one of estimating $\boldsymbol{\beta}$ in the presence of the unknown nuisance function $\lambda_0(t)$. The most widely studied of such semi-parametric structures is the proportional hazards model of Cox (1972), in which $\lambda(t) = \lambda_0(t) \exp\{\mathbf{x}(t)\boldsymbol{\beta}\}$.

An important generalization is to consider several background rate functions $\lambda_s(t)$, one for each of S strata ($s = 1, \ldots, S$). The strata may also be time-dependent, and we denote by $s(t)$ the stratum at which the subject finds himself at time t. The exposure variables are generally assumed to act in the same way (e.g., additively, multiplicatively) on each of the background rates, regardless of stratum, and a single set of β parameters is used to describe their effects. Further generalizations are possible to situations in which the background rates vary continuously with two or more continuous time variables, but these methods have not yet been fully developed and are not presented here.

(a) *Choice of basic time variable*

Substantial flexibility is available with the continuous variable models, since different choices can be made for the basic time variable t. Candidates for t include time on study, time since first employment, age and calendar year. Once t has been specified, its effects on the background mortality rates are estimated nonparametrically in $\lambda_0(t)$. The effects of the remaining time-dependent factors are then modelled in regression variables $x(t)$. Stratification of the sample into several subgroups, each with its own background mortality rate function, allows even greater flexibility. The choice of t is important, and the investigator will usually want to think carefully about the goals of the analysis before deciding which time variable to model nonparametrically and which to account for by means of regression coefficients.

Several of the analyses we have carried out have used t = age as the fundamental time variable. Secular trends in the age-specific background rates are accommodated by stratification of the sample into five- or ten-year intervals of calendar year or birthdate. One rationale for this choice of t is the fact that age is generally the most critical determinant of cancer rates. This suggests that one allow the greatest possible flexibility in their age dependence. The effects of various exposure indices that change with time on study are accommodated in the regression variables.

In other examples, particularly those involving external standard rates or in which the background age-specific rates are known to have a simple parametric form, we have examined the evolution of excess or relative risk as a nonparametric function of t = time since first exposure. Sometimes, one may wish to conduct several parallel analyses with different choices of t, in order to determine the most appropriate parametric form for each one prior to its inclusion in subsequent analyses as a time-dependent regression variable. However, some caution must be exercised in order that an inappropriate choice for t not obscure the very effects that one is looking for. For example, suppose that major attention is focused on a cumulative exposure variable $x(t)$ that is highly correlated with time on study. If time on study is selected as

t, some of the effects that rightfully should be quantified in the regression coefficient of the exposure variable will instead be hidden in the estimate of the baseline risk function $\lambda_0(t)$.

(b) *Construction of exposure functions*

One potential advantage of continuous variable methods is their ability, at least in principle, to make full use of the time history of exposures that may be recorded for each individual in the study. Estimates of annual exposure increments may be available from periodic readings of radiation dosimeters, from personal records on dates of transfer between job sites, or periodic examination of blood, urine or tissue specimens. A wide variety of exposure functions may be constructed from such data.

We first considered an approach that is of interest primarily for historical reasons. Suppose $z(u)du$ denotes the increment in exposure estimated to occur in the time or age interval $(u, u + du)$. Several investigators have constructed regression variables representing time-weighted cumulative or average exposures in the form

$$x(t) = \int_{t_0}^{t} z(u)w(t-u)\,du, \tag{5.1}$$

where t_0 is the age at entry to the study and $w(u)$ is a suitable weight function. If $w(u) = 0$ or 1, according to whether $u \leq L$ or $u > L$ years, $x(t)$ represents a lagged cumulative exposure such that increments received during the preceding L years have no effect on risk (e.g., Gilbert & Marks, 1979). By defining $w(u) = \min(1, u/L)$, exposures may be phased in linearly over a period of L years before taking maximum effect. Berry *et al.* (1979) set $w(u) = \{1 - \exp(-\lambda u)\}/\lambda$ as a method of time-weighting accumulating exposure to asbestos fibres that allows for their elimination from the lungs at rate λ. By taking $w(u)$ to be a probability density, one can express the concept of a biological latent interval as the random duration of time between an exposure increment and its effect on disease (Knox, 1973). A typical choice for w is the density function of a log-normal distribution, with mode and variance possibly estimated from the data. The 'working level month' used in the study of Rocky Mountain uranium miners is defined in precisely this way (Lundin *et al.*, 1979). We explore this method in §5.5, using data from the Montana smelter workers study.

One cause for concern regarding the uncritical use of cumulative exposure measurements is that they may fail to separate intensity and duration of exposure adequately. For example, radiation risks are commonly assessed in terms of lifetime excess cancer cases per cGy of exposure per million population, without consideration of dose fractionation or timing (Committee on the Biological Effects of Ionizing Radiation, 1980). While this practice may have some empirical justification in radiation carcinogenesis, its widespread adoption in other situations is surely to be deplored. Consider, for example, the lung cancer risk at age 60 among two smokers – one who consumed 10 cigarettes per day since age 20 and the other 20 cigarettes per day since age 40. The total number of cigarettes is the same, namely 20 pack-years or $20 \times 20 \times 365 = 146\,000$ cigarettes. However, data from the British doctors study and elsewhere suggest that the lung cancer risk is approximately proportional to $dt^{4.5}$,

where d is the number of cigarettes smoked per day and t is years of smoking (Doll & Peto, 1978; see also Example 4.7 above). This suggests that the 20-pack-year smoker who started at age 20 has $0.5(2)^{4.5} = 11.3$ times the lung cancer risk of the 20-pack-year smoker who started at age 40. Analysing the two individuals in the same category of 'cumulative dose' would be a serious error.

The choice of exposure variables used in continuous data analyses can have a major influence on the results and interpretation and on any quantitative risk assessments that are made. It is important, therefore, to demonstrate the goodness-of-fit of the resulting model and to evaluate its sensitivity to perturbations in the weight function or model equation. Even when analysing data using continuously varying baseline age rates, it is often prudent as a first step in the analysis to define the regression variables so that they represent discrete levels of intensity and duration of exposure, just as was done for grouped data. Examination of the results of such descriptive analyses can then suggest a possible role for a more quantitative approach.

We suggest that initial explorations of the data be conducted using categorical binary variables that represent different levels of each of the factors of primary interest: age at onset of exposure; intensity of exposure averaged over the period of accumulation; duration of exposure; fractionation; and time since last exposure. Trends in excess or relative risk measures according to each of these factors are of inherent interest and may help to elucidate possible underlying mechanisms. In Chapter 6, we consider how such descriptive analyses may be interpreted in terms of mathematical models of carcinogenesis. Some authors (Thomas, D.C., 1983; Brown & Chu, 1987) have successfully fitted biomathematical models directly to data from cohort studies, but, often, the quality of the data does not warrant the considerable effort that must be made to achieve a good fit, nor can competing models be clearly differentiated in terms of the weight of evidence to support them.

(c) *Some model equations*

Models are available to express the effect of regression variables $\mathbf{x}(t)$ on background rates $\lambda_0(t)$ that parallel those for the grouped data analyses considered in Chapter 4. Thus, one has

$$\lambda(t) = \lambda_0(t) + \mathbf{x}(t)\boldsymbol{\beta} \tag{5.2}$$

for an additive effect and

$$\lambda(t) = \lambda_0(t) \exp\{\mathbf{x}(t)\boldsymbol{\beta}\} \tag{5.3}$$

for a multiplicative one with multiplicative combination of risk variables (Cox, 1972). More general relative risk models may be written

$$\lambda(t) = \lambda_0(t) r\{\mathbf{x}(t)\boldsymbol{\beta}\}, \tag{5.4}$$

where, for example

$$\log r(z) = \frac{(1+z)^\rho - 1}{\rho},$$

as in (4.24). This yields the multiplicative model (5.3) at $\rho = 1$, whereas the additive relative risk model

$$\lambda(t) = \lambda_0(t)\{1 + \mathbf{x}(t)\boldsymbol{\beta}\} \tag{5.5}$$

occurs in the limit as ρ tends to 0. More general relative risk functions $r\{\mathbf{x}(t); \boldsymbol{\beta}\}$ may be constructed in which the explanatory variables $\mathbf{x}$ and the parameters $\boldsymbol{\beta}$ do not combine in the usual linear regression fashion. Frome (1983) considers models of the form

$$\lambda(t) = \lambda_0(t)\{1 + \exp(\beta_2 + x\beta_3)\}$$

for his analysis of grouped data on lung cancer and smoking from the British doctors study. He also estimates the baseline rates as a parametric function

$$\lambda_0(t) = e^{\beta_0} t^{\beta_1},$$

rather than leaving them unspecified, as suggested here.

An alternative to (5.2), in which the excess risk is a multiplicative function of the covariables, is given by

$$\lambda(t) = \lambda_0(t) + \exp\{\mathbf{x}(t)\boldsymbol{\beta}\}. \tag{5.6}$$

Pierce and Preston (1984) consider parametric models such that $\lambda(t)$ is expressed as a sum of products of linear and multiplicative terms,

$$\lambda(t) = \mathbf{x}_1(t)\boldsymbol{\beta}_1 e^{\mathbf{y}_1(t)\boldsymbol{\gamma}_1} + \mathbf{x}_2(t)\boldsymbol{\beta}_2 e^{\mathbf{y}_2(t)\boldsymbol{\gamma}_2} + \cdots,$$

where the explanatory and regression variables are partitioned $\mathbf{x} = (\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2, \ldots)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\gamma}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_2, \ldots)$. This includes (5.6) as a special case, provided that $\lambda_0(t)$ is modelled parametrically. They implement a similar generalization of the relative risk model (5.5).

(d) *External standard rates*

External standard rates are incorporated into each of these model equations just as they were for grouped data. One simply replaces the unknown functions $\lambda_0(t)$ in (5.2)–(5.6) by the quantity $\theta\lambda^*(t)$ where $\lambda^*(t)$ is the standard background rate at time t for a subject, depending upon his age and the calendar period, and $\theta = \exp(\alpha)$ is an unknown scale parameter used to adjust the standard rates so as to give the best fit to the background rates actually observed for the cohort. (Alternatively, especially in the context of (5.2), $\lambda_0(t)$ might be replaced by $\theta + \lambda^*(t)$ whereby an additive constant is used to adjust standard to background.) The known rates $\lambda^*(t)$ are typically obtained from national vital statistics, but occasionally they may come from theoretical models of the disease process, as in Example 4.7. Explicit equations for models analogous to (5.2), (5.3) and (5.5) are thus

$$\lambda(t) = \theta\lambda^*(t) + \mathbf{x}(t)\boldsymbol{\beta}, \tag{5.7}$$

$$\lambda(t) = \lambda^*(t) \exp\{\alpha + \mathbf{x}(t)\boldsymbol{\beta}\} \tag{5.8}$$

and

$$\lambda(t) = \theta\lambda^*(t)\{1 + \mathbf{x}(t)\boldsymbol{\beta}\}. \tag{5.9}$$

The continuous time version of the excess mortality ratio model considered in §4.10 is

$$\lambda(t) = \theta\lambda^*(t) + \exp\{\mathbf{x}(t)\boldsymbol{\beta}\}. \tag{5.10}$$

The availability of information on background rates is of particular importance when estimating excess risks using (5.7) or (5.10). We are unaware of any method that may exist for fitting (5.6) when $\lambda_0(t)$ is left completely unspecified. Such models may be fitted when the data are grouped, as we saw in the last chapter, or when $\lambda_0(t)$ is expressed in terms of a small number of unknown parameters, as suggested by Pierce and Preston (1984). A fully nonparametric treatment of background rates is currently limited to the multiplicative models.

In addition to the disease rates $\lambda(t)$ of primary interest, an important conceptual role is played in the sequel by a function $\nu(t)$ that represents the instantaneous rate at which subjects are lost from view during the study. Such loss may be caused by death due to 'competing' illnesses, emigration from the study area, or other reasons. We make the important assumption that ν does not depend on $\boldsymbol{\beta}$, meaning that the timing and nature of deaths due to other diseases or the withdrawal of persons from the study carry no information on how exposures affect the disease of interest. The fitted statistical model represents a 'smoothing' of the observed variation in disease rates according to exposure and other explanatory variables, in the presence of competing causes of death. Conclusions about exposure effects apply only to the conditions that prevail in the particular study and should not be expected *a priori* to hold in a population subject to other types of intercurrent mortality (Prentice *et al.*, 1978). The question as to whether or not the results can be generalized must be argued on a broader basis. These caveats apply equally, of course, to results obtained with more elementary methods.

5.2 Likelihood inference

Just as was true for grouped data analyses, statistical inference about the parameters of interest in models for continuous data requires construction of an appropriate likelihood function. Denote by t_i the age (or time) at which the ith subject ends the study, and define δ_i as 1 or 0 according to whether death or diagnosis has or has not occurred at t_i. Also denote $Y_i(t) = 1$ or 0 according to whether he is or is not under observation at age t, and let $t_i^0 = \inf\{t\colon Y_i(t) = 1\}$ denote the age at entry. General considerations suggest that the contribution of the ith subject to the likelihood function is

$$\lambda_i^{\delta_i}(t_i)\nu_i^{1-\delta_i}(t_i)\exp\left\{-\int Y_i(u)\{\lambda_i(u) + \nu_i(u)\}\,du\right\}, \tag{5.11}$$

where subscripts i have been added to the rate functions λ and ν defined earlier to indicate that they usually vary from one subject to another. The exponential term represents the probability of being disease-free between ages t_i^0 and t_i. For subjects

who develop the disease of interest ($\delta_i = 1$), the leading term $\lambda_i(t_i)$ represents the conditional probability of death or diagnosis at t_i, given that it has not occurred earlier; for those who do not ($\delta_i = 0$), the leading term $\nu_i(t_i)$ represents the conditional probability of loss. A rigorous derivation of this result requires consideration of the product integral of the instantaneous probabilities of death or disease at each age, conditional on past history (Kalbfleisch & Prentice, 1980; Johansen, 1981). Since ν_i is assumed to be free of $\boldsymbol{\beta}$, its contribution to the likelihood factor is usually ignored.

The only unknowns for models that incorporate standard rates are the scalar $\theta = \exp(\alpha)$ and the vector $\boldsymbol{\beta}$. In this case, the log-likelihood function for the entire set of cohort data may be written

$$L(\alpha, \boldsymbol{\beta}) = \sum_i \left\{\delta_i \log \lambda_i(t_i; \alpha, \boldsymbol{\beta}) - \int Y_i(u)\lambda_i(u; \alpha, \boldsymbol{\beta})\,du\right\}, \tag{5.12}$$

where $\lambda_i(t; \alpha, \boldsymbol{\beta})$ is specified by any one of the equations (5.7)–(5.10) or an analogous model formula. Formal proofs that maximum likelihood estimates based on this expression have the usual properties of consistency and asymptotic normality, with covariances estimable from the inverse information matrix, may be based on the large sample theory of counting processes (Borgan, 1984). Likelihood analyses based on (5.12) have been implemented for the multiplicative model (5.8) by Breslow *et al.* (1983), who approximate the integrals and their first and second derivatives by a summation in which the time-dependent covariables are evaluated annually for each subject. Some results of these analyses are presented in §5.5.

(a) *Poisson models for grouped data*

A formal justification for the Poisson model (4.7) used for grouped data analysis is obtained by specializing (5.11) to discrete time. Suppose that there are $J \times K$ cells or states and that $\lambda_i(t) = \lambda_{jk}$ if the ith subject is in state (j, k) at time t. This condition holds for grouped data models, in which the background rates are given by $\lambda_i(t) = \lambda_j$ and the regression variables by $\mathbf{x}_i(t) = \mathbf{x}_{jk}$ for subjects in state (j, k) at that time. Summing up the log-likelihood contributions from (5.11) over all subjects in the study leads to the total log-likelihood (Holford, 1980),

$$\sum_j \sum_k \{d_{jk} \log(\lambda_{jk}) - n_{jk}\lambda_{jk}\},$$

where d_{jk} are the numbers of deaths that occur while a subject is in the (j, k)th state and n_{jk} is the total observation time (person-years) in that state. This is precisely the log-likelihood for the Poisson distribution on which the statistical methods of Chapter 4 were based, and each of the models (5.2)–(5.10) reduces to its discrete counterpart as considered there.

(b) *Partial likelihood for multiplicative models*

The likelihood for models (5.2)–(5.6) involves the unknown nuisance functions $\lambda_0(t)$, the presence of which considerably complicates estimation of $\boldsymbol{\beta}$. Cox (1972, 1975) solved this problem for the subgroup of multiplicative models (5.3)–(5.5) by

constructing an appropriate 'partial likelihood', in which the contribution of the nuisance function is eliminated and only β remains. His approach is easily generalized to accommodate several background nuisance functions $\lambda_s(t)$ in a stratified analysis. Suppose, for example, that the ith individual is known to have died (or been diagnosed) at age t_i in calendar period (stratum) s_i. Denote by R_i the set of all subjects 'at risk' of death at that same age and period, meaning those who were alive and under observation just prior to age t_i and who were in calendar period s_i at that age. The conditional probability that the ith subject died, given that one death occurred among those in the risk set R_i, is thus

$$\lambda_i(t_i) \Big/ \Big\{ \sum_{j \in R_i} \lambda_j(t_i) \Big\}.$$

Summing up the logarithms of such contributions for all subjects who die or develop disease yields the log partial likelihood

$$L(\boldsymbol{\beta}) = \sum_i \Big[\log r\{\mathbf{x}_i(t_i); \boldsymbol{\beta}\} - \log \sum_{j \in R_i} r\{\mathbf{x}_j(t_i); \boldsymbol{\beta}\} \Big], \tag{5.13}$$

where r denotes the relative risk function. If several deaths (or disease cases) occur in a given risk set R_i, each one contributes a term to (5.13). The expression then serves as an approximation to the log partial likelihood, which is adequate so long as the deaths form only a small fraction (e.g., under 5%) of the total number in each risk set (Peto, R., 1972; Breslow, 1974).

Other methods are needed when the times of death or diagnosis are grouped, so that a substantial number d_i of cases occurs among those in the risk set at a specified t_i. Cox (1972) also proposed a linear logistic model for discrete survival data, such that each risk set yields a partial likelihood contribution proportional to

$$\frac{\prod_{j=1}^{d_i} r\{\mathbf{x}_{ij}(t_i); \boldsymbol{\beta}\}}{\sum_{\mathbf{l}} \prod_{j=1}^{d_i} r\{\mathbf{x}_{l_j}(t_i); \boldsymbol{\beta}\}}, \tag{5.14}$$

where the numerator is a product of relative risks over the d_i cases, and $\mathbf{x}_{ij}(t_i)$ denotes the covariable vector for the jth member of R_i. Assuming that the risk set also contains g_i noncases, the denominator is a summation over all $n_i \cdot Cd_i$ ways of selecting a 'control sample' containing d_i of the $n_i = d_i + g_i$ individuals in R_i. Each control sample is specified by a set of indices $\mathbf{l} = \{l_1, l_2, \ldots, l_{d_i}\}$ chosen from the numbers $\{1, 2, \ldots, n_i\}$ that identify the 'risk set' members. The labels $\{1, 2, \ldots, d_i\}$ are assumed to correspond to cases. Although the large number of terms in the denominator sum renders its calculation unfeasible if both d_i and g_i are large, recursive algorithms developed by Gail *et al.* (1981) and Storer *et al.* (1983) permit this approach to be used when there is a moderate number of cases – say, no more than 20 or so – in each risk set.

For the special case $r(\mathbf{x}; \boldsymbol{\beta}) = \exp(\mathbf{x}\boldsymbol{\beta})$, Andersen and Gill (1982) show that the usual likelihood calculations based on differentiation of (5.13) yield asymptotically normal estimates, the variances and covariances of which may be estimated from the observed

information matrix. Prentice and Self (1983) derived analogous results for models (5.4) and (5.5), in which the relative risk function is given by $r(\mathbf{x}; \boldsymbol{\beta}) = 1 + \mathbf{x}\boldsymbol{\beta}$ or a more general expression. As already mentioned, satisfactory nonparametric methods of estimation for the additive models (5.2 and 5.6) have not yet been developed.

Example 5.1

Hutchinson and colleagues (1980) conducted a historical cohort study of nearly 1500 women treated for benign breast disease to determine their subsequent incidence of breast cancer. A later analysis of these data related each woman's history of treatment with hormones (oestrogens) to breast cancer risk (Thomas, D.B. *et al.*, 1982). The data used here for illustrative purposes were compiled by Persing (1981) from 1353 cases with a histological confirmation of the initial benign lesion.

A simple tabulation of the data, shown in Table 5.1, leads to a relative risk estimate for hormone users *versus* nonusers of $(25 \times 499)/(33 \times 522) = 0.72$, and suggests a possible protective effect of the oestrogens. However, it ignores the person-years of observation denominators and, more importantly, the relationship between the age at which each woman started to take oestrogens and the age at which she developed, or was at risk of developing, breast cancer. Oestrogen use and age were strongly related, since the cohort had been assembled during a 35-year surgical practice over which time there were marked changes in the use of oestrogens for contraception or treatment of post-menopausal symptoms.

A partial likelihood analysis was undertaken with t = age in order to account for the age dependence of both exposure and disease risk. Table 5.2 lists the integral ages t_i at which diagnoses of breast cancer were made and the composition of the risk sets for 1036 women for whom it was known whether or not and, if so, at what age, oestrogen use began. A woman contributed to the risk set R_i provided that her benign breast disease had been diagnosed before age t_i, so that she was under observation, and provided also that she had not yet died, been lost to follow-up or otherwise removed from risk of breast cancer, for example, by having a double mastectomy.

In this example, there is a single, age-dependent binary covariate $x_1(t)$ indicating whether or not a woman has received oestrogen. It is defined as 1 for all women in R_i who received hormone prior to t_i, i.e., for the women in columns labelled H1 in Table 5.2, and 0 for the remaining women. Note that a woman's covariable value may change from $x_1(t) = 0$ to $x_1(t) = 1$ as she is followed forward in the study. A parallel analysis in terms of a fixed (i.e., not age-dependent) covariable, taking values 1 or 0 according to whether a woman ever received oestrogen (columns H1 and H2), yields fallacious results, since some women are then analysed as 'exposed' at ages before the exposure actually began.

Suppose that the relative risk function is defined by $r(x; \beta) = \exp(x\beta)$, so that the relative risk is $\psi = \exp(\beta_1)$ for prior exposure $\{x_1(t) = 1\}$ and $1 = \exp(0)$ for no prior exposure $\{x_1(t) = 0\}$. The data in each risk set are conveniently arranged in a 2×2 table of exposed *versus* nonexposed and cases *versus*

Table 5.1 Distribution of 1353 women treated for benign breast disease according to history of oestrogen use and development of breast cancer[a]

Oestrogen use	Breast cancer		
	Yes	No	Total
Yes	25	522	547
No	33	499	532
Unknown	8	266	274
Total	66	1287	1353

[a] From Persing (1981), from data originally collected by Hutchinson *et al.* (1980)

Table 5.2 Composition of the risk sets at each age of diagnosis of breast cancer

Age = t_i	Total number in risk set R_i	Cancer cases[a]			Non-cancer cases[a]		
		H1	H2[b]	no H	H1	H2	no H
30	148	0	0	1	27	63	57
37	279	0	0	1	38	130	110
38	304	1	0	2	40	139	122
41	409	0	0	2	58	180	169
44	507	1	0	2	99	200	205
45	528	2	0	2	109	194	221
46	567	1	0	3	135	192	236
48	602	1	0	3	154	179	265
49	602	0	0	1	178	160	263
50	610	2	0	1	196	140	271
51	598	0	0	3	216	115	264
52	577	2	0	2	226	94	253
54	520	1	0	1	221	66	231
58	389	4	0	0	158	35	192
60	348	2	0	0	147	17	182
61	313	1	0	1	124	14	173
62	285	1	0	1	100	11	172
64	234	2	0	1	73	9	149
65	200	1	0	0	55	7	137
67	159	0	0	1	40	2	116
68	137	1	0	1	29	2	104
69	121	0	0	3	27	0	91
76	37	0	0	1	5	0	31

[a] H1, hormone (oestrogen) users at ages less than or equal to t_i; H2, hormone users at ages greater than t_i; no H, hormone nonusers

[b] This column contains only zeros, since women who developed breast cancer at age t_i were removed from further study

noncases. For example, at age $t_i = 52$ we have

	Exposed	Nonexposed	
Cases	2 (e_i)	2	4 (d_i)
Noncases	226 ($f_i - e_i$)	347	573 (g_i)
Total	228 (f_i)	349	577 (n_i)

The contribution to the numerator of the partial likelihood* (5.14) for the risk set at age $t_i = 52$ is thus $\psi^2 1^2 = \psi^2$. More generally, if e_i of the d_i cases are exposed, the contribution is ψ^{e_i}. If a 'control' sample $\{l_1, \ldots, l_{d_i}\}$ in the denominator yields u exposed and $n_i - u$ nonexposed, its contribution to the denominator is ψ^u. Since the number of such samples with exactly u exposed is

$$\binom{d_i}{u}\binom{g_i}{f_i - u},$$

the total contribution of the risk set R_i to the partial likelihood is proportional to

$$\frac{\binom{d_i}{e_i}\binom{g_i}{f_i - e_i}\psi^{e_i}}{\sum_{u=0}^{d_i}\binom{d_i}{u}\binom{g_i}{f_i - u}\psi^u}. \tag{5.15}$$

For example, the risk set at age $t_i = 52$ years yields the contribution

$$\frac{\binom{4}{2}\binom{573}{226}\psi^2}{\binom{4}{0}\binom{573}{228}\psi^0 + \binom{4}{1}\binom{573}{227}\psi^1 + \binom{4}{2}\binom{573}{226}\psi^2 + \binom{4}{3}\binom{573}{225}\psi^3 + \binom{4}{4}\binom{573}{224}\psi^4}.$$

For this special case of a single binary exposure variable, the partial likelihood (5.15) is identical to the exact conditional likelihood used for estimation of relative risk in a series of 2×2 tables compiled from case-control data (§7.5 Volume 1; Breslow, 1976). The computer program LOGODDS, presented in Appendix VI of Volume 1, may be utilized for this special problem, although some modifications are needed to accommodate the large binomial coefficients that occur in (5.15).

The full partial likelihood is a product of terms of the form (5.15), one for each line in Table 5.2. Maximizing this, we find $\hat{\psi} = 1.80$. The Mantel–Haenszel test of the hypothesis H_0: $\psi = 1$ (§4.4, Volume 1), known also as the logrank test (Peto, R. & Peto, 1972), yields $\chi^2 = 4.41$ ($p = 0.02$). We conclude that oestrogen use significantly increased the breast cancer risk in this population of women with benign breast disease. However, part of the observed association might be related to the confounding effects of other risk factors that were increasing with calendar time. Both oestrogen use and breast cancer incidence were rising during the course of the study, and inclusion of year of birth as an additional covariate in the model reduced the estimated relative risk for oestrogen to $\hat{\psi} = 1.49$, $\chi^2 = 1.82$ (Thomas, D.B. *et al.*, 1982).

Repeating the analysis in terms of the improper (fixed) exposure covariate yields $\hat{\psi} = 0.70$, $\chi^2 = 1.59$ (NS), a result rather close to that for the summary data in Table 5.1 where we ignored age altogether. Careful examination of Table 5.2 shows the reason for the discrepancy. When averaged over the 23 risk sets, with weights proportional to their size, the proportion of women who used oestrogen at any time (H1 + H2) is 0.38 for cases and 0.52 for noncases. However, the average proportions of women who had started using oestrogens previously are instead 0.38 and 0.29. In other words, when cases and noncases are compared in terms of whether or not they had a history of exposure *at the same age*, the cases are more likely to have used the hormone. More noncases were observed during the later ages and calendar periods, at which oestrogen treatment was more common.

We tested whether or not the relative risk for oestrogen use varied with age by including an age-dependent covariable $x_2(t) = x_1(t) \times (t - 55)$ in the model $\lambda(t) = \lambda_0(t) \exp\{\beta_1 x_1(t) + \beta_2 x_2(t)\}$. Here, $\psi = \exp(\beta_1)$ denotes the relative risk at age 55, while $\exp\{\beta_2(t - 55)\}$ is a multiplicative factor that measures the change in the relative risk for younger or older women. Alternatively, we could have set $x_2(t) = x_1(t) \times \log(t/55)$, in which case the relative risk would be modelled as a power function $\psi(t/55)^{\beta_2}$. With the addition of $x_2(t)$ to the model, the contributions to (5.15) become

$$\frac{\binom{n_i}{e_i}\binom{g_i}{f_i - e}\exp[e_i\{\beta_1 + \beta_2(t_i - 55)\}]}{\sum_{u=0}^{d_i}\binom{n_i}{u}\binom{g_i}{f_i - u}\exp[u\{\beta_1 + \beta_2(t_i - 55)\}]}.$$

Using once again a modification of the program LOGODDS, we find $\hat{\beta}_1 = 0.614 \pm 0.285$, $\hat{\beta}_2 = 0.017 \pm 0.029$ and a score statistic for testing $\beta_2 = 0$ of $\chi_1^2 = 0.32$ (NS). Thus, there is no evidence for a trend in the relative risk with age.

An explicit formula for the score statistic used to test $\beta_2 = 0$ was given in Volume 1, equation (4.31). Contrary to the assertion made there, however, the estimates $\hat{\psi}$ of relative risk inserted in equations 4.30 and 4.31 of Volume 1 must be maximum (partial) likelihood estimates in order that these statistics have asymptotic chi-square distributions under the null hypothesis. Modifications of the equations are needed

when the Mantel-Haenszel estimate $\hat{\psi}_{MH}$ is used in place of the maximum likelihood estimate. (See Breslow *et al.* (1984) for the modification needed in the test for trend, and Tarone (1985) for the corresponding global test of homogeneity of relative risk.)

(c) *Goodness-of-fit*

The goodness-of-fit of models fitted to grouped cohort data may be evaluated relatively easily by comparing the observed and fitted numbers of deaths in each cell of the cross-classification, by plotting the adjusted residuals (4.13) in various ways and by examining the summary chi-square (4.6) or deviance (4.12) goodness-of-fit statistics. Indeed, an advantage of this approach is that one is almost forced to examine how well the model predicts the outcome in each cell. Unfortunately, no such safeguard is built into the continuous data analysis, and extra steps are needed to determine whether or not the model provides a reasonable summary of the observed data.

One of the most important methods for examining the goodness-of-fit of the proportional hazards model was introduced in Example 5.1. It involves adding to the model age- or time-dependent covariables that represent the interaction of exposure effects with those of age or time. Such covariables typically take the form $y(t) = x(t)\log(t/c)$ or $y(t) = x(t)(t-c)$, where c is a constant representing a standard age and $x = x(t)$ represents an exposure that may or may not be time-dependent. The sign of the regression coefficient estimated for $y(t)$ indicates whether the trend in relative risk associated with a given amount of exposure is increasing or decreasing with age. Additional interaction variables with quadratic terms $(t-c)^2$ or $\log^2(t/c)$ may be needed if the relative risk first rises and then declines with age.

An alternative approach that may be implemented without explicit recourse to age-dependent covariables is to carry out separate analyses for each of two or three age intervals by dividing the risk sets into groups depending on t_i. Comparison of regression coefficients for the same exposure variables in different age groups indicates the direction of any trend, and comparison of the maximized partial likelihood for the combined analysis with the sum of the maximized partial likelihoods for the separate analyses provides a formal test of the statistical significance of the differences in the coefficients.

A third approach that retains some of the features of the grouped data analysis is to define a partition of age into J intervals and exposure into K categories. Separate binary covariables are then defined for each of the JK cells in the cross-classification. The score test for the addition of these covariables to the regression models compares the observed and expected numbers of cases in each cell. However, since the expected values are based on the model fitted to continuous data, the simple $\sum(O-E)^2/E$ chi-square formula does not apply (Schoenfeld, 1980; Tsiatis, 1980). It is necessary to estimate the covariances of the $O-E$ differences in order to carry out the test.

A graphical approach to the evaluation of goodness-of-fit of proportional hazards is to partition the sample into a small number of (possibly time-dependent) categories of persons with similar exposure histories. Separate estimates of the age-specific disease incidence functions are modelled for each one. When plotted against age on a semilogarithmic scale, these curves should stay roughly a constant distance apart if the

hypothesis of proportionality holds. This procedure is illustrated below with the benign breast disease data. (See especially Figure 5.2.)

A variation of this graphical analysis is helpful when the exposure variables are numerous, and estimation of a separate age incidence function within categories of exposure is not feasible. One defines a partition of the data into K subgroups on the basis of the estimated relative risk function $r\{\mathbf{x}(t); \hat{\boldsymbol{\beta}}\}$. If $\mathbf{x} = \mathbf{x}(t)$ depends on age, therefore, so will subgroup membership. Separate estimates of the age-incidence curves for each subgroup, say, $\hat{\lambda}_k(t)$ for $k = 1, \ldots, K$, are compared with the fitted age-incidence curves $r_k\hat{\lambda}_0(t)$, where r_k is the average relative risk in the kth subgroup and $\hat{\lambda}_0(t)$ is the background age incidence function estimated from the total cohort. Breslow (1979) gives an illustration using data from clinical trials.

The addition of exposure × age interaction variables to the basic equation is also applicable as a means of assessing goodness-of-fit when the background rates are assumed to be proportional to external standard rates or are modelled parametrically. A graphical method for evaluating the proportionality assumption is illustrated in Figures 5.4 to 5.6.

(d) *Nonmultiplicative models*

Partial likelihood unfortunately provides only a partial solution to the problem of fitting continuous models to cohort data. The approach is not applicable if the basic model is additive, for example, or has any other form in which the exposure effects do not act multiplicatively on the background rates. It is necessary in such circumstances to assume that the background rates are given by some formula that depends on parameters α and to base the inference on the general log-likelihood (5.12). This is precisely what one does when the background rates are assumed to be known up to a constant $\theta = \exp(\alpha)$ of proportionality, or when explicit parameters are used to represent background rates by age and year in grouped data analyses.

(e) *Notes on computing*

Example 5.1 is a very special case in that it involves only a single binary covariable. This allows the data to be represented as a series of 2 × 2 tables and allows use of programs for the regression analysis of log odds ratios in 2 × 2 tables in the analysis. Most problems, including analyses of the data on Montana smelter and on Welsh nickel workers, presented below, involve multiple discrete and continuous regression variables $\mathbf{x}(t)$. Here, the computing problems are considerably more complex. One must either compute and store the covariable history $\mathbf{x}(t)$ for each individual at times $t = t_i$ for *each* of the risk sets R_i in which he appears, or else supply a set of covariable function subroutines that calculate the requisite covariables, at different times, from basic data available for each subject. An exception is the additive relative risk model (5.5), for which only the covariable values for the cases and the *average* of the covariables for the other risk-set members need to be stored (Gilbert, 1983; Prentice & Mason, 1986). For large cohort studies, it is generally not possible to store all the data needed in the central memory of a computer. This means that a separate pass through

the data files is made at each iteration of the procedure used to find the maximum partial likelihood estimate $\hat{\beta}$. The program must also be capable of accommodating time-dependent stratification, whereby the stratum index for each subject is available from stored data, or from function subprogram calculations, for each risk set in which he appears.[1]

5.3 Nonparametric estimation of background rates

Nonparametric estimates of cumulative disease incidence or death rates based on continuous data sampled from a homogeneous population were introduced in Volume 1 (§2.3) with an illustrative application to data on mouse skin tumours. Virtually identical techniques are used to estimate cumulative disease rates from cohort data. Suppose that the distinct times or ages at which deaths or cases occur are $0<t_1<t_2<\cdots<t_I$. Denote by d_i the number of cases at t_i and by $n_i = d_i + g_i$ the total size of the risk set R_i, i.e., the number of cohort members under observation at t_i. Let $\Lambda(t) = \int_0^t \lambda(s)\,ds$ denote the unknown cumulative rate in the general population. The usual estimate of Λ, often ascribed to Nelson (1969), is

$$\hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}. \tag{5.16}$$

Some motivation for this formula is provided by the fact that the differentials

$$\frac{\hat{\Lambda}(t_i) - \hat{\Lambda}(t_{i-1})}{t_i - t_{i-1}} = \frac{d_i}{n_i(t_i - t_{i-1})},$$

which equal the observed number of deaths divided by the approximate person-years observation time in the age interval (t_{i-1}, t_i), are obvious estimates of the corresponding instantaneous rates.

The variance of $\hat{\Lambda}(t)$ is estimated using Greenwood's (1926) formula

$$\operatorname{Var} \hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{5.17}$$

This is the continuous data analogue of equation (2.2) for the standard error of a cumulative or directly standardized rate calculated from grouped data. When considered as a random function of t, $\hat{\Lambda}$ is approximately distributed as a Gaussian stochastic process with mean $\Lambda(t)$ and a covariance function $C(s, t) = \operatorname{Cov}\{\hat{\Lambda}(t), \hat{\Lambda}(s)\}$ that is estimated for $t \leq s$ by (5.17) (Breslow & Crowley, 1974). This fact has enabled statisticians to develop simultanous confidence bands for $\Lambda(t)$, or the corresponding 'survival' function $S(t) = \exp\{-\Lambda(t)\}$, over an interval of time or age (Gillespie & Fisher, 1979; Hall & Wellner, 1980).

The same approach may be used to obtain separate estimates of cumulative hazard

[1] Pat Marek of the Fred Hutchinson Cancer Research Center (see Peterson *et al.*, 1983) developed the program that we used for the illustrative analyses presented here. This program is currently being simplified and adapted to run on microcomputers.

or mortality within each of several subsets or strata. One simply classifies the d_i deaths and n_i risk-set members according to the particular stratum in which they appear at time t_i. As we noted earlier, plots of the estimated $\hat{\Lambda}_s(t)$ for different strata are useful for examining the consistency of the data with the assumption of proportional hazards. If the disease incidence rates $\lambda_1(t)$ and $\lambda_2(t)$ are in constant ratio $\lambda_2(t) = \theta\lambda_1(t)$, then so are the integrated hazards $\Lambda_2(t) = \theta\Lambda_1(t)$. Plots of $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ on a semilogarithmic scale should therefore be roughly a constant distance apart.

Example 5.2

From the data in Table 5.2 and equation (5.16), one may construct an estimate of cumulative breast cancer incidence for women who had no prior exposure to oestrogen and another for women with such exposure. For example, the cumulative incidence at age $t = 45$ for women without prior exposure is estimated to be

$$\hat{\Lambda}_1(45) = \frac{1}{121} + \frac{1}{241} + \frac{2}{263} + \frac{2}{351} + \frac{2}{407} + \frac{2}{417} = 0.0354,$$

whereas for women with an exposure history it is

$$\hat{\Lambda}_2(45) = \frac{1}{41} + \frac{1}{100} + \frac{2}{111} = 0.0524.$$

Figure 5.1 shows these two functions, plotted using arithmetic (Fig. 5.1A) and logarithmic (Fig. 5.1B) scales for $\hat{\Lambda}$. Although there is considerable instability in the estimates due to the small numbers, there is no evidence of a systematic trend in the difference between the two curves on the semilogarithmic plot. This confirms the results of the formal analysis of Example 5.1 in which we tested whether the ratio of rates for exposed *versus* unexposed showed a trend with age and concluded that the assumption of proportionality was justified.

Note that the estimated lifetime cumulative incidence for oestrogen nonusers in this cohort is approximately twice that of the general population rate of 7%. The rates for users are even higher. This illustrates the fact that a history of benign breast disease itself augments the subsequent breast cancer risk (Hutchinson *et al.*, 1980).

(a) *Smoothed estimates of age- or time-specific rates*

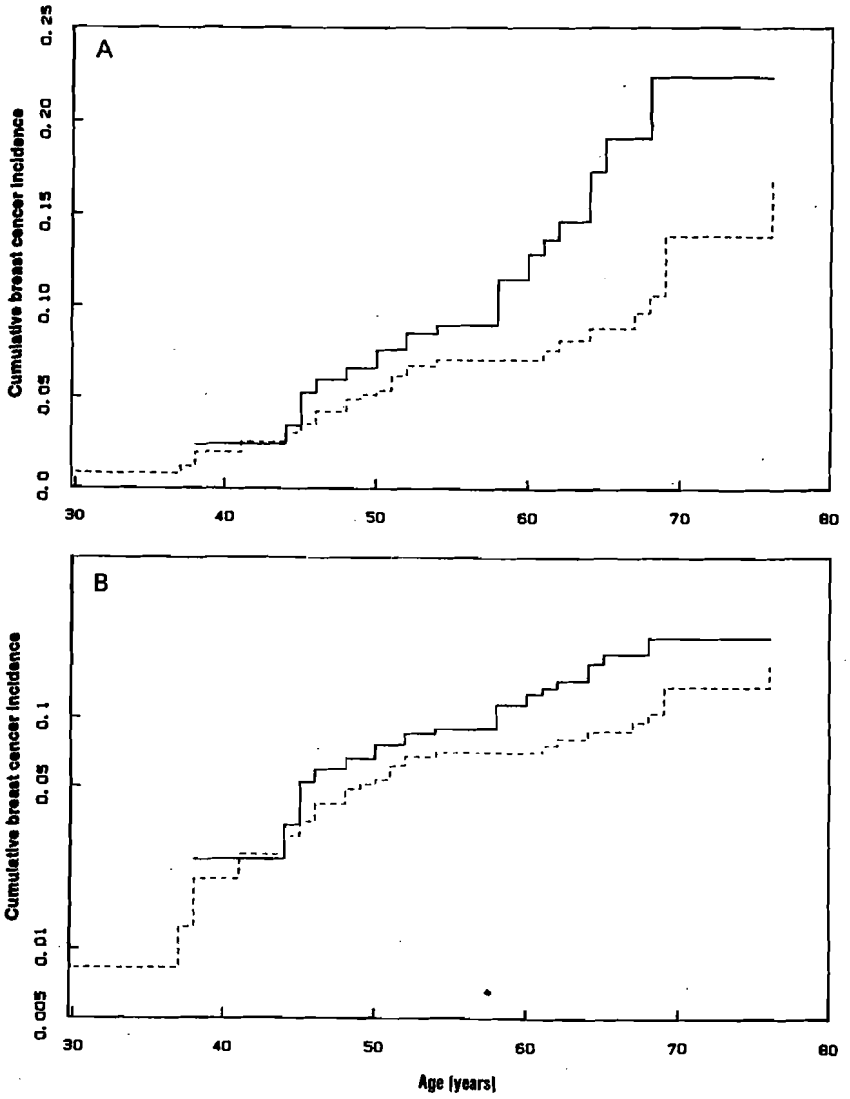
Estimates of cumulative incidence or mortality rates such as shown in Figure 5.1 are not as informative as they might appear at first sight. They tend to overemphasize the jumps that occur at very high ages, at which the estimate is least stable due to declining numbers at risk. Also, the age- or time-specific rates are usually of greater intrinsic interest than the cumulative rate. Recent work by Ramlau-Hansen (1983) and Yandell (1983) has validated kernel estimates of $\lambda(t)$ that have the form

$$\hat{\lambda}(t) = \frac{1}{b}\int_0^\infty K\left(\frac{t-s}{b}\right) d\hat{\Lambda}(s) = \frac{1}{b}\sum_{i=1}^{I} K\left(\frac{t-t_i}{b}\right)\frac{d_i}{n_i}. \tag{5.18}$$

Here, $K(x)$ is a smooth, positive kernel function integrating to one, and b is a bandwidth that determines the degree of smoothness in the estimate. Thus, $\hat{\lambda}(t)$ is simply a weighted average of the increments d_i/n_i in $\hat{\Lambda}(t)$, with K defining the weights and b the size of the 'window' about t within which the estimates of the instantaneous rates are averaged. Its standard error is given by

$$\operatorname{SE}\{\hat{\lambda}(t)\} = \frac{1}{b}\left\{\sum_{i=1}^{I} K^2\left(\frac{t-t_i}{b}\right)\frac{d_i}{n_i^2}\right\}^{1/2}. \tag{5.19}$$

Fig. 5.1 Cumulative incidence of breast cancer for women with benign breast disease with (solid line) and without (dotted line) prior exposure to oestrogen. (A) Arithmetic scale; (B) log scale



In the examples below we have used the kernel defined by $K(x) = (0.75)(1-x^2)$ for $-1 \leqslant x \leqslant 1$, and $K(x) = 0$ elsewhere. Bandwidths are varied to achieve a compromise between too much random noise (small b) and too great a loss of structure in the estimated rates (large b). The final choice is based largely on visual appearance, although objective criteria are also available (Titterington, 1985). Note that $\hat{\lambda}(t)$ is defined only over the interval $(t_1 + b,\ t_I - b)$, where t_1 and t_I are the minimum and maximum times at which cases were observed to occur. In a refinement of this method, Tanner and Wong (1984) select the bandwidths depending on age, so that they are narrow when deaths are frequent and risk-set sizes are large, and wide elsewhere.

Example 5.3

Figure 5.2 graphs smoothed estimates of breast cancer incidence for the data on women with benign breast disease shown in Table 5.2. These were obtained from the cumulative incidence estimates $\hat{\Lambda}$ shown in Figure 5.1 by applying (5.18) with $K(x) = 0.75(1-x^2)$ for $|x| \leqslant 1$ and two bandwidths $b = 10$ (Fig. 5.2A) and $b = 15$ years (Fig. 5.2B). Relatively large bandwidths were necessary to achieve statistical stability because of the small number of cases in this study, namely 23 among women with prior exposure to oestrogens and 34 among those not so exposed. Consequently, they may obscure somewhat the true variation in incidence with age. Note the greater degree of smoothing achieved with the larger bandwidth. Although the rate ratio for exposed *versus* unexposed seems to increase slightly over the 40–65-year age range, we already know from the partial likelihood analysis in Example 5.1 that this trend is not statistically significant.

The observation that the age-specific rates are nearly constant over the age range shown, especially for women with no prior exposure to oestrogen, is not surprising. As mentioned in the previous example, there was a strong birth cohort effect on the age-specific breast cancer rates in this particular population. Since the data are analysed here on a cross-sectional basis, ignoring birth cohort, the observed age-incidence curve is distorted (flattened) in comparison with the more typical pattern of rising incidence until the age of menopause with a change in slope thereafter. A similar phenomenon was observed in Volume 1 for breast cancer rates in Iceland that were analysed according to both calendar year and birth cohort. Compare Figures 2.3 and 2.4 in Volume 1, and also Figure 4.2.

(*b*) *Estimating baseline rates under the multiplicative model*

These techniques are easily extended to provide estimates of the cumulative baseline rate function

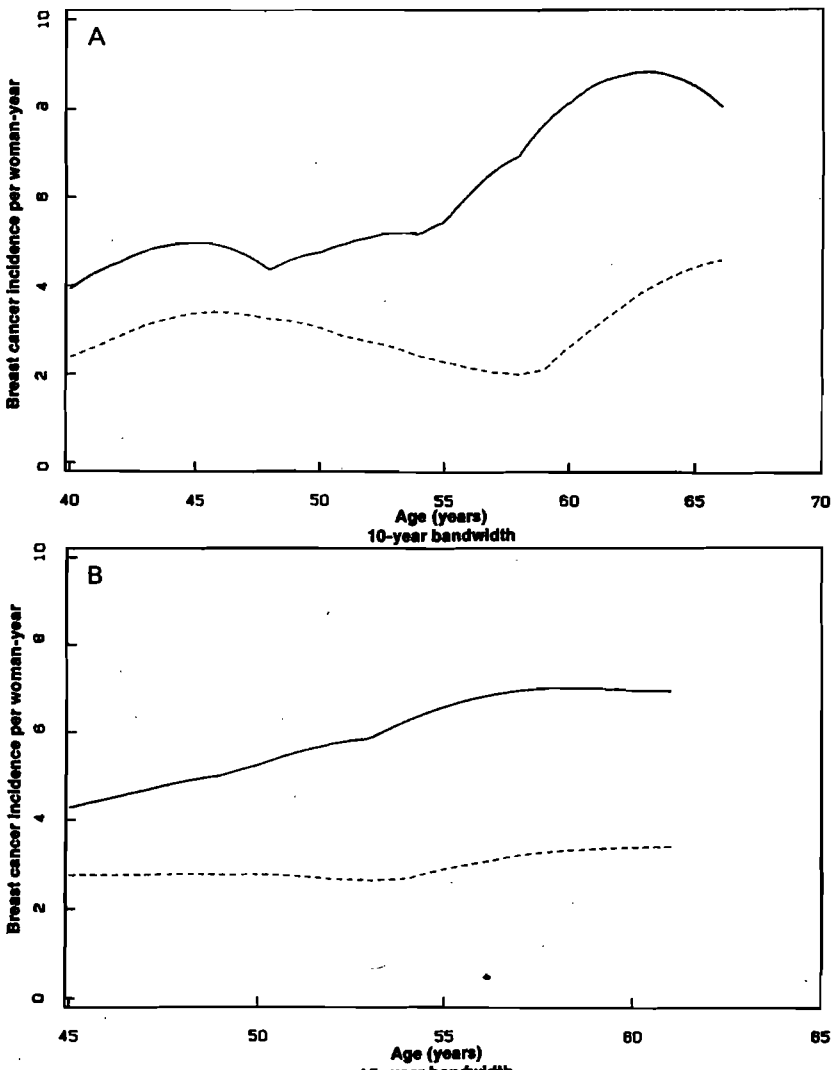
$$\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$$

under the various multiplicative models proposed for heterogeneous samples. Using a heuristic argument to achieve joint maximum likelihood estimation of Λ_0 and $\boldsymbol{\beta}$ in Cox's (1972) model (5.3), Breslow (1974) derived the estimate

$$\hat{\Lambda}_0(t) = \sum_{t_i \leqslant t} \frac{d_i}{\sum_{j \in R_i} \exp\{\mathbf{x}_j(t_i)\hat{\boldsymbol{\beta}}\}}, \tag{5.20}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum partial likelihood estimate from (5.13) or (5.14). The obvious

Fig. 5.2 Smoothed estimates of breast cancer incidence for women with benign breast disease with (solid line) and without (dotted line) prior exposure to oestrogen. (A) Ten-year bandwidth; (B) 15-year bandwidth



extension for the general multiplicative model is to

$$\hat{\Lambda}_0(t) = \sum_{t_i \le t} \frac{d_i}{\sum_{j \in R_i} r\{\mathbf{x}_j(t_i); \hat{\boldsymbol{\beta}}\}}. \tag{5.21}$$

The main difference between (5.20) or (5.21) and the equation applicable to homogeneous samples is that the size of the risk set at t_i, which appears in the denominator of (5.16), is replaced by the total estimated relative risk for the risk set at that time. Tsiatis (1981) has shown that $\hat{\Lambda}_0(t)$ defined by (5.20) also has an asymptotic Gaussian distribution.

If the data are stratified, separate estimates of the background rates

$$\Lambda_s(t) = \int_0^t \lambda_s(u)\, du$$

are obtained for each stratum simply by restricting the deaths and risk sets in (5.20) or (5.21) to that stratum. Smoothed estimates of the age-specific baseline rates $\lambda_s(t)$ are available *via* (5.18). However, their standard errors are more complicated than that shown in (5.19) because of the need to account for the error in estimation of $\boldsymbol{\beta}$ (Andersen & Rasmussen, 1982). We present some illustrative examples in §§5.5 and 5.6 below.

(c) *Nonparametric estimation of relative mortality functions*

An extension of the multiplicative models, incorporating external standard rates, allows the equations and computer programs already developed for nonparametric estimation of cumulative baseline rates to be used also for nonparametric estimation of cumulative *relative* mortality functions (Andersen *et al.*, 1985). Consider first the simple model $\lambda(t) = \theta\lambda^*(t)$, whereby each subject's disease rate is assumed to be equal to a constant multiple of the standard rate for a person of the same age and sex. Maximization of the parametric likelihood (5.12) in this situation yields the usual ratio of observed to expected deaths, i.e., the SMR

$$\hat{\theta} = \frac{\sum_{i=1}^{I} d_i}{\sum_{i=1}^{I} \int Y_i(u)\lambda_i^*(u)\, du}, \tag{5.22}$$

as the 'optimal' estimate (Breslow, 1975). Here, Y_i is as defined in §5.2.

One way of looking for changes in the SMR that would invalidate its use as a single summary measure is to divide the age or time axis into a number of discrete intervals and to cumulate the deaths and integrated standard rates within each one. The methods developed for testing the homogeneity of such SMRs with grouped data (§3.4) continue to apply and indeed are strongly recommended. Formal justification is provided in terms of a generalization of the basic model to $\lambda(t) = \theta_k\lambda^*(t)$ for $t_{k-1} < t \le t_k$.

A further generalization of this approach allows the SMR to be modelled as a continuous function of time, i.e., $\theta(t) = \lambda(t)/\lambda^*(t)$ or $\lambda(t) = \theta(t)\lambda^*(t)$. Comparing this

formula with (5.3), we note that the two models are formally identical: $\theta(t)$ plays the role of the unknown baseline rate $\lambda_0(t)$, and $\log \lambda^*(t)$ is a time-dependent covariate with known regression coefficient $\beta = 1$. Just as we were earlier able to estimate the cumulative baseline rate

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$$

nonparametrically in terms of a step-function (equations 5.16, 5.20 and 5.21), here we are able to estimate the cumulative or integrated SMR

$$\Theta(t) = \int_0^t \theta(u)\, du.$$

Note that the cumulative SMR equals the average SMR over the time interval $(0, t)$ multiplied by the length of the interval. It is measured in units of time. $\Lambda(t)$, however, is the product of a rate with time and is thus dimensionless. These differences notwithstanding, an estimate of the integrated SMR is obtained from (5.20) as

$$\hat{\Theta}(t) = \sum_{t_i \le t} \frac{d_i}{\sum_{j \in R_i} \lambda_j^*(t_i)}. \tag{5.23}$$

The estimate of the average SMR over the time interval (t_{i-1}, t_i) is thus given by the number of deaths or cases observed at time t_i divided by the total expected number among the risk-set members.

Introduction of explanatory variables $\mathbf{x}(t)$ into the model allows covariance adjustment of the nonparametric SMR estimates. In its most general form, the underlying model for the unknown disease rate is written

$$\lambda(t) = \theta(t)\lambda^*(t) r\{\mathbf{x}(t); \boldsymbol{\beta}\}.$$

The β parameters in the relative risk function are estimated by a generalization of the partial likelihood (5.14), namely

$$\prod_{i=1}^{I} \frac{\prod_{j=1}^{d_i} \lambda_j^*(t_i) r\{\mathbf{x}_{ij}(t_i); \boldsymbol{\beta}\}}{\sum_{l} \prod_{j=1}^{d_i} \lambda_{lj}^*(t_i) r\{x_{lj}(t_i); \boldsymbol{\beta}\}}. \tag{5.24}$$

In practice, $\lambda_j^*(t_i)$ or its logarithm is incorporated into the model as an 'offset' or covariable with known regression coefficient. Once $\boldsymbol{\beta}$ is obtained *via* maximization of (5.24), the integrated SMR is estimated as

$$\hat{\Theta}(t) = \sum_{t_i \le t} \frac{d_i}{\sum_{j \in R_i} \lambda_j^*(t_i) r\{\mathbf{x}_{ij}(t_i); \hat{\boldsymbol{\beta}}\}}, \tag{5.25}$$

generalizing (5.21). Adjusted or unadjusted estimates $\hat{\Theta}(t)$ based on (5.25) and (5.23), respectively, are smoothed *via* the kernel method to yield nonparametric estimates of the SMR:

$$\hat{\theta}(t) = \frac{1}{b} \int_0^\infty K\left(\frac{t-s}{b}\right) d\hat{\Theta}(s) = \frac{1}{b} \sum_{i=1}^{I} K\left(\frac{t-t_i}{b}\right) \frac{d_i}{R_i^+}, \tag{5.26}$$

where R_i^+ is the total standard risk at t_i, either

$$R_i^+ = \sum_{j \in R_i} \lambda_j^*(t_i)$$

for the unadjusted estimate or

$$R_i^+ = \sum_{j \in R_i} \lambda_j^*(t_i) r\{\mathbf{x}_{ij}(t_i); \hat{\boldsymbol{\beta}}\}$$

for the adjusted one. The standard error of the unadjusted estimate is

$$\text{SE}\, \hat{\theta}(t) = \frac{1}{b}\left[\sum_{i=1}^{I} K^2\left(\frac{t-t_i}{b}\right) \frac{d_i}{(R_i^+)^2}\right]^{1/2}, \tag{5.27}$$

analogous to (5.19).

Sections 5.5 and 5.6 contain several illustrations of nonparametric estimation of baseline and relative disease mortality functions and the fitting of multiplicative models to continuous cohort data by partial likelihood. Flexible model structures are available even within the multiplicative environment by varying the fundamental time variable t, the definitions of the covariables $\mathbf{x}(t)$ and the relative risk function r. The choice should be made separately for each study, taking into account the goals of the investigation and the nature of the available data. If good a-priori information suggests that the background rates are of a simple parametric form, it may be preferable to model them by time-dependent covariables, rather than nonparametrically in the function $\lambda_0(t)$. For example, population data and multistage theory both suggest that cancer incidence rates are proportional to a power of age. Defining one of the covariables $x(t)$ to be the logarithm of age at 'time' t, this sort of age dependence is easily accommodated in relative risk functions of the form $r\{\mathbf{x}(t); \boldsymbol{\beta}\} = \exp\{\mathbf{x}(t)\boldsymbol{\beta}\}$. In this case, and also when the background rates are assumed to be proportional to standard rates $\lambda^*(t)$, one may want to set t = 'time since onset of exposure' in order to have a nonparametric evaluation of the evolution of relative risk with continuing exposure. Alternatively, if we set t = age and incorporate $x(t) = \log(t)$ into the exponential relative risk function, our nonparametric estimate of $\theta(t)$ *via* (5.26) provides a graphical evaluation of the goodness-of-fit of the assumed parametric model.[1]

5.4 Sampling from the risk sets

Implementation of the methods of analysis of continuous data outlined in the preceding sections is expensive and time-consuming in the case of data from large cohort studies. This is true whether one uses external standard rates and the log-likelihood (5.12) or adopts the partial likelihood approach based on (5.13) or (5.14). In the former instance, the basic data for each subject are needed to re-evaluate integrals of the form $\int Y_i(u)\lambda_i(u; \alpha, \boldsymbol{\beta})\, du$ at each cycle of iteration. In the latter case, one must re-evaluate the relative risks $r\{\mathbf{x}_j(t_i); \boldsymbol{\beta}\}$ for each subject in every risk set in

[1] Recent work by F. O'Sullivan at the University of California, Berkeley, on spline-smoothed hazard estimates with cross-validation may offer some advantages over the kernel methods suggested here.

which he appears (except, as noted earlier, for the additive relative risk model). It is often possible to store some intermediate quantities, such as the covariable values $\mathbf{x}_j(t_i)$ for each subject at each time of death, for use in subsequent iterations. However, this may not be advisable if it greatly increases the amount of reading the computer does from disk files.

(*a*) *Complexity of partial likelihood analyses*

Suppose that the basic time variable is in fact age and that birth cohort or calendar year is accounted for by stratification. Let R denote the risk set consisting of all persons being followed in the study at a given age t at some time during a specified calendar period s. In practice, we have found that integral ages and five- or ten-year calendar periods generally provide sufficient accuracy for construction of the risk sets. Because of ties in the recorded data, several deaths may occur in some of the risk sets. This would not happen if they were defined in terms of exact (continuous) ages at death. However, since the number of deaths or cases is generally much smaller than the total size of the risk set, which may well be of the order of hundreds or even thousands depending on the size of the original cohort, the approximation inherent in the use of (5.13) with such tied data is excellent.

Example 5.4

Table 5.3 shows the distribution by age and calendar period of 142 respiratory cancer deaths that occurred among the Montana smelter workers during the years 1938–1963, this being the period of follow-up of the initial study reported by Lee and Fraumeni (1969). When classified by integral age at death and by calendar year in six intervals of five years or less, they define 91 separate risk sets. Most risk sets contain a single respiratory cancer death, but the multiplicities range as high as $d_i = 4$, for example, among workers aged 51 or 67 during the period 1955–1959. Also shown for each risk set are the numbers of deaths from other causes, these being the matched 'controls' one would use in a proportional mortality analysis.

Table 5.4 presents the numbers of noncases (g_i) for each of the risk sets defined in Table 5.3. These range from 17 workers (in addition to the one case) under observation at age 84 during 1950–1954, to 880 workers on study at age 40 during 1955–1959. The mean risk-set size was 322, with a standard deviation of 215. Thus, each of the 8014 subjects appeared on average in 3.6 risk sets. Since the calculations needed for a partial likelihood analysis treat each such risk-set appearance as a separate observation, the effective 'sample size' is of the order of 30 000 observations, of which 142 are cases. This gives some feeling for the magnitude of the computing problem.

(*b*) *Risk-set sampling*

It is evident from equations (5.13) and (5.14) that the information about relative risks associated with the exposure variables is provided by a comparison of the exposures of the case(s) with the exposures of the remainder of the cohort members in each risk set. Since most risk sets are very large in comparison with the number of cases, little information would be lost if the comparison were made between the cases and a small sample of 'controls' drawn randomly from among the other cohort members in the risk set. This is the idea of matched 'case-control within a cohort' sampling proposed by Thomas, D.C. (1977) for efficient analyses of continuous cohort data. Mantel (1973) earlier suggested a similar strategy for stratified analyses under the label 'synthetic retrospective study'. As emphasized in Volume 1, the idea of sampling controls from an on-going but unobserved (and possibly only conceptual) cohort

Table 5.3 Respiratory cancer deaths (d) and deaths from other causes ($t-d$) for the Montana cohort by age and year; construction of the risk sets[a]

Age (years)	Calendar year											
	1938–1939		1940–1944		1945–1949		1950–1954		1955–1959		1960–1963	
	d	$t-d$	d	$t-d$	d	$t-d$	d	$t-d$	d	$t-d$	d	$t-d$
40									1	4		
45			1	2								
46					3	5					1	6
47							1	5				
48			1	5								
49							1	5	2	3		
50			1	0					1	10	1	4
51									4	10		
52					1	3			1	6		
53			2	3			1	16	1	13	1	10
54							2	6	2	13		
55									3	11	1	6
56											2	9
57			1	5	1	8	1	7	3	15	1	14
58			1	7	3	4			1	13	4	12
59					1	3					1	13
60							2	6	3	5	1	12
61							2	8	2	7	3	11
62			2	5			1	7	2	11	3	11
63			1	6	1	10			1	12	3	11
64			1	6	1	5			1	12		
65					1	4	1	4	2	9	2	9
66	2	1	1	2			1	11			1	10
67			1	3	1	10	1	9	4	7	1	8
68			1	4			3	6	1	9	3	5
69			1	5			1	11	1	9	1	3
70							2	9	1	9	1	8
71	1	1									1	10
72											1	7
73					1	8	1	7	1	13	1	6
74					1	7			3	16		
76					2	8			3	10		
79							1	4				
80											1	6
81											1	4
83							3	4				
84							1	1				

[a] Entries appear for a given age/calendar year only if one or more respiratory cancer deaths occurred.

Table 5.4 Numbers of Montana smelter workers alive and under observation at particular ages and calendar periods; sizes of the risk sets with cases excluded

Age (years)	Calendar year					
	1938–1939	1940–1944	1945–1949	1950–1954	1955–1959	1960–1963
40					880	
45		344				
46			504			762
47				688		
48		309				
49				644	745	
50		279			726	722
51					726	
52			374		722	
53		252		516	707	645
54				484	663	
55					635	587
56						607
57		241	289	374	538	583
58		258	264		484	569
59			270			511
60				290	398	487
61				284	358	441
62		200		266	337	400
63		167	254		312	371
64		151	238		282	
65			209	240	248	299
66	56	143		229		261
67		146	168	208	217	246
68		137		211	195	230
69		125		193	201	195
70				167	192	184
71	37					163
72						144
73			102	102	163	136
74			85		138	
76			61		87	
79				46		
80						53
81						37
83				25		
84				17		

investigation is one of the main justifications for the validity of inferences made in actual case-control investigations.

(c) *Likelihood analysis*

Under the general multiplicative model, the contribution to the likelihood from a risk set containing d cases with exposure variables $\mathbf{x}_1(t), \ldots, \mathbf{x}_d(t)$, and m randomly sampled 'controls' with exposure variables $\mathbf{x}_{d+1}(t), \ldots, \mathbf{x}_{d+m}$, is proportional to

(Prentice & Breslow, 1978)

$$\frac{\prod_{j=1}^{d} r(\mathbf{x}_j(t); \boldsymbol{\beta})}{\sum_{\ell} \prod_{j=1}^{d} r(\mathbf{x}_{\ell_j}(t); \boldsymbol{\beta})}. \tag{5.28}$$

The numerator of this expression is the product of the relative risks for the actual cases. The denominator summation is over all possible subsets $\ell = (\ell_1, \ldots, \ell_d)$ of size d drawn from the $d+m$ members of the risk set, there being $\binom{d+m}{d}$ such subsets in all. Each may be thought of as representing a possible set of d cases that might have been observed to die from the cause of interest at time t and whose relative risk product is compared to that for the actual cases. Precisely the same likelihood is used for the matched analysis of actual case-control studies. However, in Volume 1 (equation 7.1), we restricted consideration to multiplicative relative risk functions of the form $r(\mathbf{x}; \boldsymbol{\beta}) = \exp(\mathbf{x}\boldsymbol{\beta})$. The same expression is used also for the full partial likelihood analysis (5.14), except that there the denominator sum is taken over the much larger number $\binom{n}{d}$ of subsamples drawn from the full risk set.

An important feature of the case-control within a cohort method of analysis is that the time-dependent covariables for the controls need be evaluated only at the particular age t for which they are sampled. Once calculated, they are easily stored in a rectangular data array in central memory for efficient computer processing. In a partial likelihood analysis, the time-dependent covariables for each cohort member must usually be re-evaluated for each risk set in which he appears.

(d) *Model selection and regression diagnostics*

The primary advantage of the risk-set sampling methodology is that it reduces the effective number of observations to a reasonable size for efficient computer processing. This encourages the investigator sitting at a computer terminal to fit a variety of models involving different exposure variables to the sampled data and select those that fit well for further examination. Such interactive data analysis is often not possible with a full partial likelihood approach. Depending on the size of the data set and the available computer, one may have to wait several hours or even overnight before seeing the results of a particular fit.

Regression diagnostics for matched case-control and partial likelihood analyses, analogous to those considered in §4.3 for grouped data, have recently become available as a result of work by Pregibon (1984), Moolgavkar *et al.* (1984) and Storer and Crowley (1985). For the most part, these are developed in terms of approximate changes in estimated regression coefficients or test statistics that would accompany deletion of individual observations (cases or controls), or deletion of entire risk sets. As shown earlier, such diagnostics are helpful in evaluating the stability of the fitted model and the extent to which the results depend on data for only one or a few individuals. An illustration of their use in case-control within a cohort analyses appears in §5.6. One may also use the predicted within-risk set 'probability of being a case' as a

guide to goodness-of-fit. This is defined for each subject as his estimated relative risk divided by the sum of relative risks for the entire risk set (assuming one case per set). Such predicted probabilities are usefully summed across individuals when there are particular covariate values for comparison with the corresponding observed numbers. They may also be used to define 'residuals' for case-control studies.

(e) *Estimating background and relative rates from the case-control samples*

An examination of equations (5.20), (5.21), (5.23) and (5.25), used to estimate the cumulative background rates $\Lambda(t)$ or the cumulative relative rates $\Theta(t)$, suggests how they may be adapted to serve also for case-control samples. The essential requirement is that one know the sampling fractions used to select controls within each risk set, i.e., the total size of the risk set from which the controls are sampled. This requirement is met for the 'synthetic' case-control technique suggested here, where one explicitly constructs the risk sets using the cohort data base and then carries out the control sampling by computer. It usually will not be met for case-control studies conducted outside the context of a cohort study. One then needs supplementary data in order to estimate absolute risks.

Denote by $\mu_i = \mu_i(\hat{\beta})$ the average of the estimated relative risk factors associated with the $n_i = d_i + g_i$ subjects in the ith risk set. Thus,

$$\mu_i = \frac{1}{n_i} \sum_{j \in R_i} r_{ij},$$

where r_{ij} is the estimated relative risk $r\{\mathbf{x}_j(t_i); \hat{\beta}\}$, or estimated absolute risk $\lambda_j^*(t_i) r\{\mathbf{x}_j(t_i); \hat{\beta}\}$, depending upon whether Λ or Θ is under consideration. The estimates $\hat{\Lambda}$ and $\hat{\Theta}$ may both be expressed in the general form

$$\sum_{t_i \le t} (d_i / n_i \mu_i).$$

If we lack data for the entire risk set but do have available a sample of m_i controls drawn without replacement from the g_i noncases in R_i, we could estimate μ_i by the sample mean

$$\bar{r}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} r_{ij}.$$

A refinement would be to substitute $n_i^{-1}\{d_i \bar{s}_i + g_i \bar{r}_i\}$ for $\bar{r}_i$, where $\bar{s}_i$ denotes the average (relative) risk for the d_i cases. However, this should make little difference unless the cases constitute a large fraction of the risk set. Substituting $\bar{r}_i$ for μ_i in (5.21) thus yields

$$\hat{\Lambda}_0(t) = \sum_{t_i \le t} \frac{d_i}{n_i \bar{r}_i} \tag{5.29}$$

as our approximation to $\hat{\Lambda}$, and a similar substitution in (5.25) gives an approximation to $\hat{\Theta}$.

The main drawback to this approach is the fact that the reciprocal of a sample mean is a biased estimator of the mean. The problem is acute for the small control sample

sizes typically used, with m_i in the range from 1 to 20. Breslow and Langholz (1987) suggest two possible ways of correcting the bias in (5.29) to yield a better estimate. The most promising, based on a Taylor series expansion of $\bar{r}^{-1}$ about μ^{-1}, leads to the equation

$$\hat{\Lambda}_T(t) = \sum_{t_i \le t} \frac{d_i}{n_i \bar{r}_i} \left\{ 1 - \frac{\hat{\sigma}_i^2}{m_i \bar{r}_i^2} \right\}, \tag{5.30}$$

where $\hat{\sigma}_i^2 = (m_i - 1)^{-1} \sum_j (r_{ij} - \bar{r}_i)^2$ is the within-risk-set variance. The other, derived from the jackknife principle of Quenouille (1949), leads to

$$\hat{\Lambda}_J(t) = \sum_{t_i \le t} \frac{d_i}{n_i \bar{r}_i} \left\{ m_i - \frac{(m_i - 1)^2}{m_i} \sum_{j=1}^{m_i} \frac{1}{(m_i - r_{ij}/\bar{r}_i)} \right\}. \tag{5.31}$$

Note that (5.30) and (5.31) both reduce to (5.29) if $r_{ij} = \bar{r}_i$ for all of the sampled controls.

Section 5.5 illustrates the application of these equations to data from the Montana cohort (see especially Figure 5.8). Neither applies very well for $m_i = 5$, but both perform satisfactorily for $m_i = 20$. If only five controls or fewer are available from each risk set, it is probably wise to pool the controls sampled from each R_i with those from neighbouring risk sets R_j, i.e., those with $|t_j - t_i| < b$ where b is a designated bandwidth, in order to increase the effective number of controls for each. The rationale for this procedure is that the average (relative) risk μ_i should be reasonably constant over risk sets within a narrow time interval, since their membership will change little.

(f) *Selection of controls*

The procedure recommended here for construction of the matched sets of cases and controls that will actually be used in the analysis is as follows: First select from the risk set R_i all d_i cases that develop or die from the disease of interest at time t_i. Then select m_i controls, *at random and without replacement*, from among the g_i members of R_i who do *not* develop the disease at that time. The total of d_i cases and m_i sampled controls then constitutes a reduced risk set R_i^*.

Early theoretical arguments given in support of this procedure (Prentice & Breslow, 1978) assumed that the number g_i of potential controls was effectively infinite. This meant that there would be no overlap between the controls sampled from different risk sets, nor would a subject who later developed disease be sampled as a control. In practice, of course, this assumption is not met. Indeed, the risks sets corresponding to advanced ages are often quite small (see Table 5.4), and it may be desirable to sample all available controls from them. It is then quite conceivable that an individual sampled as a control at one age will turn out to be a case later on or be sampled again as a control at that time. With the methodology employed here, therefore, the R_i^* can and do overlap, at least on occasion.

The fact that the reduced risk sets R_i^* may not be disjoint in finite cohorts has caused some concern about the validity of the inference procedure implicit in the use of (5.28), since this approach combines statistical information from each of them as if they were statistically independent. For example, Lubin and Gail (1984) mentioned the possibility

of excluding previously chosen controls from consideration as future controls, yet including them as cases if and when they developed the disease. If the original risk sets R_t are small, however, this latter procedure is biased (Robins *et al.*, 1986a). Prentice *et al.* (1986) propose a rather more elaborate sampling procedure in which the controls sampled along with a case from R_t are also considered as controls in *each* of the risk sets in which that case previously appeared. This increases the amount of information available in the case-control sample by increasing the sizes of the sampled risk sets. However, it also introduces correlations between the partial likelihood contributions from different risk sets which then need to be accounted for in the analysis. In order to avoid these complications, and also to keep the effective sample size small enough to permit interactive analyses, we prefer the procedure outlined above in the context of case-control analysis of assembled cohort data. Oakes (1981) and Cox and Oakes (1984, section 8.8) have shown that the product of terms (5.28) is still a partial likelihood (Cox, 1975) and that estimates and standard errors derived from them have the same asymptotic validity as those based on all the data.

The question of the number of controls that should be sampled from each risk set is considered in §7.6.

(*g*) *Computer programs*

Appendix IV of Volume 1 contained the source code for a computer program that implemented matched case-control analyses based on the conditional likelihood (5.28) with $r(\mathbf{x};\boldsymbol{\beta}) = \exp(\mathbf{x}\boldsymbol{\beta})$, $d = 1$ and variable m. Another program, given in Appendix V of Volume 1 (Smith *et al.*, 1981) permitted arbitrary numbers of cases and controls in each stratum or risk set. However, since the relative risk function was restricted to the log-linear form and since the program used an inefficient method of evaluating the denominator of (5.28) and related expressions, it is now outmoded. Gail *et al.* (1981) developed a more efficient algorithm for the log-linear model using a recursive method of calculation. This approach was developed further by Storer *et al.* (1983) so as to permit additive and other more general relative risk functions. The latest version of their program, known as PECAN, mimics the GLIM syntax for specifying terms in the model, allows for variable factoring and offsets to the regression equation, and provides an option for calculation of regression diagnostics in the manner of Storer and Crowley (1985).

5.5 Analyses of continuous data from the Montana smelter workers cohort

From descriptions of the Montana smelter workers study and the grouped data analyses presented earlier, especially in Examples 2.1 and 2.2 and in §§3.2, 4.5 and 4.8, the reader should already have a good understanding of how the occurrence of respiratory cancer in this cohort is related to date of hire, birthplace and duration of work in moderate or high arsenic exposure areas. In this section, we elaborate by reporting the results of fitting of continuous models to the original data set, consisting of 8014 individual data records containing details of exposure history and follow-up. Due to the complexity of the partial likelihood calculations, fitting each model

generally required an overnight computer run in batch mode. In spite of this effort, the results serve mostly to confirm what has already been learned from the more economical grouped data analyses. They do not provide any really new insights.

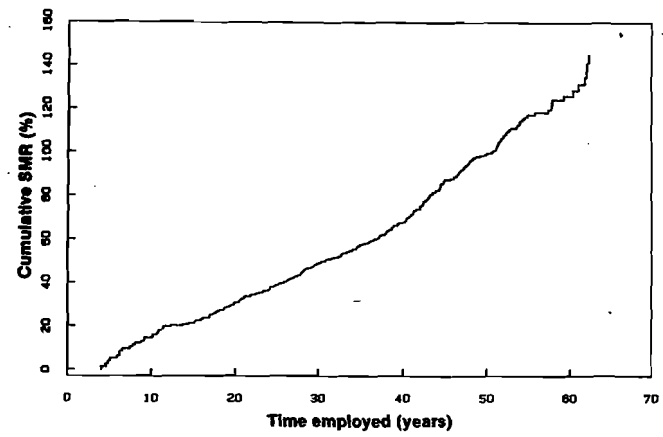
(*a*) *Respiratory cancer SMR and years since first employed*

The simple ratios of observed to expected numbers of respiratory cancer deaths shown in Table 4.17 increased markedly about 30 years or so after date of initial employment. Here, we take a more detailed look at this change in relative risk using the nonparametric estimate (5.23) of the cumulative SMR, defining t = 'years since initial employment'. Using all 288 respiratory cancer deaths, including 12 at 80 years of age or older that were excluded from most previous analyses, we obtained the results shown in Figure 5.3. The first case occurred at 4.07 years from date of hire and the last at 62.25 years. The cumulative SMR climbs steeply for the first few years, rises more gradually until about 35 years, and then steepens again. However, just as is true for estimates of the cumulative mortality function, it is hard to get a good visual impression of the SMR itself from this graph alone.

A much better representation of the temporal changes in the SMR is provided in Figure 5.4, where we graph the smoothed SMRs calculated from (5.26) using bandwidths of five and ten years. The ten-year bandwidth results in a substantially smoother curve, but also restricts the range over which the estimate is available. The sharp rise in the SMR appears to begin at about 30 years using the ten-year bandwidth and a little later with the shorter width. Such details may be obscured with a grouped analysis.

Figure 5.5 presents 90% confidence bands for the SMR estimated using a five-year

Fig. 5.3 Cumulative standardized mortality ratio (SMR) for respiratory cancer, by number of years since initial employment for Montana smelter workers



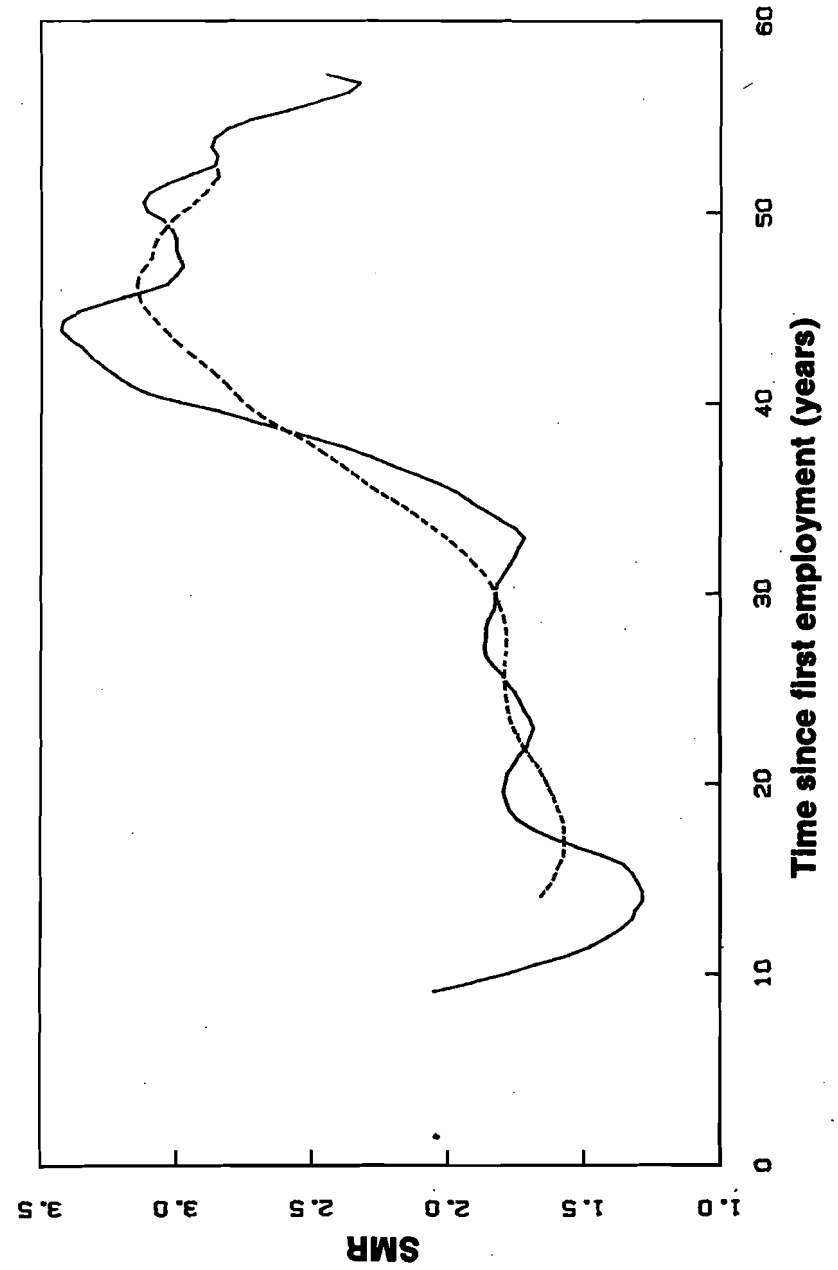


Fig. 5.4 Smoothed estimates of the standardized mortality ratio (SMR) for respiratory cancer, by years since initial employment for Montana smelter workers using five- (——) and ten-year (– – –) bandwidths

Fig. 5.5 Ninety percent confidence bands (– – –) for the smoothed standardized mortality ratio (SMR) for respiratory cancer (——), Montana smelter workers

bandwidth. The confidence bands were derived on the log scale in order to approximate more closely a normal error distribution. Specifically, we used the formula

$$\log \hat{\theta}(t) \pm 1.645 \times \{\mathrm{SE}(\hat{\theta}(t))\}/\hat{\theta}(t)$$

where $\mathrm{SE}(\hat{\theta}(t))$ is given by (5.27).

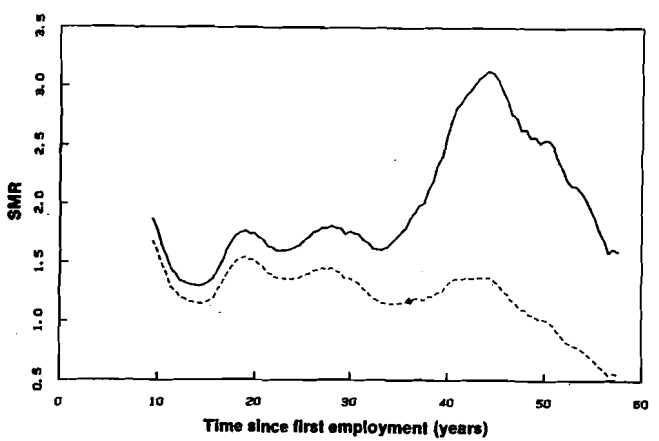
One possible interpretation of the results depicted in Figure 5.4 would be that the Montana cohort as a whole had somewhat elevated rates of respiratory cancer in comparison with the US population, perhaps because of a higher prevalence of cigarette smokers, but that the specific effects of the arsenic exposure did not become manifest until after a latent period of some 30 years. However, we already know from our analyses in Table 4.18 that this interpretation is probably fallacious. Because of the study design, namely the fact that follow-up began no earlier than 1938 whereas the first employees were hired before the turn of the century, most of the person-years of observation for those hired before 1925 occurred in the interval from 25 to 63 years from date of hire. Since we already know that the SMR for those hired before 1925 is much greater than for those hired later, it seems likely that the apparent rise at 30 years from date of hire is an artefact caused by confounding with period of hire.

In order to confirm this latter interpretation, we conducted a proportional hazards regression analysis based on equation (5.3) with t = 'years since first employment'. In addition to the log standard rates $x_0(t) = \log\{\lambda^*(t)\}$, the covariables were x_1, a binary

indicator of date hired coded 1 for 1885–1924; x_2 a binary indicator of birthplace coded 1 for foreigh born; $x_3(t)$ a lagged, continuous, time-dependent covariable giving the number of years worked in moderate arsenic exposure areas at time $t-2$; and $x_4(t)$, as for x_3 for years worked in a heavy arsenic exposure area. There were 280 distinct times at which cases occurred (for eight pairs of cases, the recorded values of years for employment to death were tied), and thus 280 separate risk sets containing as many as 5000 members each. Even on a large computer system, the partial likelihood fitting of the model with four covariables would have been prohibitively expensive and time-consuming. For this reason, we rounded time since initial employment to the midpoint of the corresponding year, and also excluded the 12 deaths that occurred at age 80 and above, thereby reducing the number of risk sets from 280 to 57 for the adjusted analysis.

The regression coefficients (± standard errors) estimated with this approach for the four covariables were: $\hat{\beta}_1 = 0.70 \pm 0.18$, $\hat{\beta}_2 = 0.47 \pm 0.14$, $\hat{\beta}_3 = 0.017 \pm 0.007$ and $\hat{\beta}_4 = 0.041 \pm 0.010$. Two smoothed estimates of the SMR were constructed using a five-year bandwidth – one with and one without covariable adjustment. The difference is striking (Fig. 5.6). The curve calculated without covariable adjustment closely resembles that in Figure 5.4 but is a bit smoother due to the fact that some averaging took place by consolidating the number of risk sets from 280 to 57. The adjusted curve has a shape that closely resembles the unadjusted one for the first 30 years, but remains roughly constant thereafter and even starts to decline to values below 1.0. The sharp peak noted in the unadjusted SMR is thus entirely explained by the four covariables and mostly, as we have previously noted, by the first one. What appears from Figure 5.4 to be evidence for a 'latent interval' turns out on closer examination to be an artefact caused by the confounding effects of year of first employment.

Fig. 5.6 Smoothed estimates of the standardized mortality ratio (SMR) for respiratory cancer, by years since first employment for Montana smelter workers, with (– – –) and without (——) adjustment for covariable effects



The adjusted curve in Figure 5.6 represents the SMR for a baseline category of US-born smelter workers hired in 1925 or after who spent their entire work history in 'light' arsenic exposure areas. If the model is reasonably correct, such workers had respiratory cancer rates that were only slightly elevated over those of the US population. There is no suggestion that the relative risk increased with time since initial employment once account is taken of the covariables. If anything, it declined!

(b) *Comparison of grouped and continuous data analyses*

Similar conclusions regarding the cohort to national rate ratio and its evolution in time may be drawn from the grouped data results presented in Table 4.19. See especially the middle column of that table, in which variations in the SMR with calendar year of follow-up (rather than time since initial exposure) are investigated. We estimated a rate ratio for US-born workers hired after 1924 with 'light' arsenic exposure of $\exp(0.581) = 1.79$ for the first calendar period of follow-up (1938–1949), but this declines to $\exp(0.581-0.480) = 1.11$ during the last period (1970–1977).

Table 5.5 compares the results of a grouped analysis of the Montana data (Table 4.19, column 3) with the results from a partial likelihood analysis of the full data set.

Table 5.5 Regression coefficients and standard errors from multiplicative models fitted to grouped and continuous data from the Montana smelter workers study: 1938–1977[a]

Regression variables	Method of analysis	
	Grouped	Continuous (partial likelihood)
All covariables binary (0/1)		
Employed before 1925	0.444 ± 0.151	0.405 ± 0.153
Foreign-born	0.445 ± 0.153	0.484 ± 0.154
Moderate arsenic[b]		
1–4 years	0.600 ± 0.166	0.601 ± 0.166
5–14 years	0.259 ± 0.242	0.261 ± 0.243
15+ years	0.684 ± 0.206	0.674 ± 0.207
Heavy arsenic[b]		
1–4 years	0.193 ± 0.305	0.170 ± 0.312
5+ years	1.069 ± 0.230	1.088 ± 0.232
Deviance	282.1	−3167.0[c]
Continuous arsenic variables		
Employed before 1925	0.441 ± 0.151	0.403 ± 0.153
Foreign-born	0.432 ± 0.153	0.473 ± 0.153
Years moderate arsenic[a] (×10)	0.222 ± 0.067	0.218 ± 0.068
Years heavy arsenic[a] (×10)	0.662 ± 0.138	0.664 ± 0.139
Deviance	292.1	−3177.0[c]

[a] From Breslow (1985a)
[b] Lagged two years
[c] Twice log-likelihood

Exposure variables for the partial likelihood analysis first were defined with discrete values that indexed the same categories of exposure that were used earlier to group the data. The results in the first part of the table indicate an excellent agreement between the two methods. This is not surprising in view of the fact that precisely the same model structures were used for relative risk. The grouped data analysis accounted for age and year effects by stratification into 16 age × year cells (four ten-year intervals for each) and explicit estimation of the corresponding parameters. The partial likelihood analysis accounted for age and year by stratification of the 276 respiratory cancer deaths into 167 risk sets on the basis of integral age at death and five-year calendar period. Evidently the age and year effects have been dealt with adequately by the broad categories used for grouping the data. There is little point in carrying out the costly and time-consuming partial likelihood analysis in this case.

The second part of Table 5.5 presents results for a partial likelihood analysis that incorporates the continuously changing arsenic variables defined by numbers of years of work in moderate or heavy exposure areas. A rather crude approximation to this continuous analysis can be obtained with the grouped data by assigning quantitative exposure values to each level of the two factors for arsenic exposure duration. From a sample consisting of 20 controls drawn from each risk set, we estimated that the average number of years of moderate arsenic exposure in the <1-year category was 0.05775 years, in the 1–4 category 2.272 years, in the 5–14 category 8.746 years and for the 15+ category 29.74 years. The corresponding averages for the three categories of heavy arsenic exposure were 0.0205, 2.219 and 16.69 years, respectively. These values were used to define the two quantitative variables for the grouped analysis. In spite of the rather approximate nature of their definition, the agreement between the grouped and continuous data analyses is still remarkably good.

Some information regarding the adequacy of the relative risk function $\exp(x\beta)$ proposed for the continuous exposure variable analyses is available by comparing the goodness-of-fit measures in the two parts of Table 5.5. Whether obtained from grouped or continuous analyses, there is a difference of 10.0 between the two measures of fit. Although the justification is approximate for the continuous analysis (due to the fact that the continuous exposure variables cannot be exactly represented as linear combinations of the corresponding discrete exposure variables), we referred this value to tables of chi-square on three degrees of freedom to gauge the relative merits of each fit and found $p = 0.02$. Thus, the assumption of a linear increase in log relative risk with increasing duration of exposure does *not* appear to be a tenable one. The separate coefficients for moderate arsenic exposure suggest that a plateau is reached after one year of exposure, whereas with heavy arsenic exposure the main effect is not seen until five or more years following exposure. See also Example 3.6.

(c) *External standard rates* versus *partial likelihood*

Breslow *et al.* (1983) conducted a partial likelihood analysis of continuous data from the Montana cohort and a parallel analysis based on the parametric likelihood (5.12) using US death rates for white males in five-year intervals of age and calendar year as a standard. These analyses, which were based on follow-up through 1963 only and ignored date of hire, are not comparable with those presented elsewhere in this monograph. The results are reproduced here because we did not wish to undertake the cumbersome job of reanalysing the 1938–1977 data using the fully parametric model. The sizes of the risk sets used in this analysis are those shown in Tables 5.3 and 5.4.

There were three exposure variables: x_1, a binary indicator of birthplace, coded 1 for foreign born; $x_2(t)$, a continuous, age-dependent variable specifying the number of years employed in one or more of the areas said to have moderate levels of arsenic exposure; and $x_3(t)$, defined analogously to x_2 for heavy arsenic exposure. The latter two variables were constructed from personnel records that allowed determination of the number of years a worker spent at moderate or high arsenic exposure levels for each of the seven calendar periods pre-1938, 1938–1939, 1940–1944, . . . , 1960–1963. The relative risk function that related these variables to the age- and year-specific background rates was $\mathrm{RR} = \exp\{\beta_1 x_1 + \beta_2 x_2(t) + \beta_3 x_3(t)\}$ for the partial likelihood and $\mathrm{RR} = \exp\{\alpha + \beta_1 x_1 + \beta_2 x_2(t) + \beta_3 x_3(t)\}$ for the parametric analysis.

The parametric analysis entailed approximation of the integral expression (5.12) and its first and second partial derivatives by a summation over years of calendar time. Functions of the covariable values evaluated at annual intervals were multiplied by each subject's contribution to the expected number of deaths (i.e., standard death rate × time on study during the year), and these products were summed over all calendar years that the subject was in the study.

The first two columns of Table 5.6 contrast the parameter estimates and standard errors obtained using these two very different approaches. There is again substantial agreement between the estimated regression coefficients. The parametric model, incorporating the external standard rates, also allows estimation of the constant term $\hat{\theta} = \exp(\hat{\alpha})$, which represents the SMR for cohort members with zero covariable values, relative to the national population. Since $\hat{\theta} = \exp(0.61) = 1.84$, one would interpret the results as saying that US-born workers who remained in light exposure

Table 5.6 Parameter estimates (± standard errors) obtained by fitting a variety of multiplicative models to continuous data from the Montana study: 1938–1963[a]

Regression variable		Method of analysis					Proportionate mortality (other deaths as controls)
		Parametric based on standard rates	Partial likelihood	Case and m controls			
				$m = 20$	$m = 10$	$m = 5$	
Constant	α	0.61 ± 0.12	—	—	—	—	—
Foreign-born	β_1	0.76 ± 0.18	0.72 ± 0.20	0.70 ± 0.21	0.66 ± 0.23	0.75 ± 0.25	0.72 ± 0.23
Moderate arsenic (×10)	β_2	0.22 ± 0.07	0.22 ± 0.07	0.21 ± 0.08	0.29 ± 0.10	0.35 ± 0.11	0.22 ± 0.10
Heavy arsenic (×10)	β_3	0.58 ± 0.13	0.60 ± 0.13	0.69 ± 0.16	0.74 ± 0.18	0.85 ± 0.23	0.53 ± 0.18

[a] From Breslow *et al.* (1983)

areas had respiratory cancer rates approximately 84% in excess of those of US white males of the same age. Foreign-born workers experienced mortality rates approximately $\exp(0.76) = 2.1$ times higher than this. For each year spent in a moderate or heavy arsenic exposure area, these rates were increased roughly by another 2% (moderate exposure) or 6% (heavy exposure). Of course, from our earlier analyses of grouped data for 1938–1977 (see especially Table 4.19), we know that these results are confounded to some extent with the effect of period of hire and that the change in relative risk with additional arsenic exposure, especially at moderate levels, does not increase smoothly as assumed by the model.

The excellent agreement between the results of the two analyses indicates that variations in the SMR by age and calendar year do not seriously confound the comparisons of SMRs for foreign- *versus* US-born or those with different degrees of arsenic exposure. Furthermore, when interaction terms were added to the partial likelihood model, there was no indication that the effects of the exposure variables changed systematically with age or year. This provides some mild evidence in support of the multiplicative model. However, with the grouped analysis of the data for 1938–1977 (Table 4.19), we noticed some confounding between calendar year of follow-up and period of hire, a variable that had been ignored in the analysis of the data for 1938–1963.

(*d*) *Efficiency gains from use of an external standard*

Perhaps just as striking as the agreement between the regression coefficients is the agreement in their standard errors as estimated by parametric and semiparametric (partial likelihood) analyses (Table 5.6, columns 1 and 2). According to the results of Oakes (1977, 1981), one would expect a substantial gain in efficiency from the use of external standard rates only if exposures varied between risk sets, that is to say with age and year. Consider a single exposure X considered as a random variable sampled from the risk sets. The relative efficiency of β estimation for the partial likelihood analysis, under the null hypothesis $\beta = 0$, is given by $E\{\mathrm{Var}(X \mid R)\}/\mathrm{Var}(X)$ where $\mathrm{Var}(X)$ denotes the total and $\mathrm{Var}(X \mid R)$ the conditional (within risk set) variance. A similar result holds for the alternative hypothesis $\beta \neq 0$, provided that the sampling probabilities for drawing subjects from risk sets are made proportional to their relative risks of death under the model.

In order to evaluate this result empirically, we estimated the within (σ_W^2) and between (σ_B^2) risk-set components of variance for each of the three exposure variables used in the analysis. We found ratios $\sigma_B^2/(\sigma_B^2 + \sigma_W^2)$ of 15.9% for birthplace, 4.5% for moderate arsenic exposure and 1.7% for heavy arsenic exposure. This is consistent with the small increases observed in estimated standard errors between parametric and partial likelihood analyses, these being about 10% for birthplace and smaller for the coefficients of the arsenic exposure duration variables.

(*e*) *Results of sampling from the risk sets*

Table 5.6 also shows the regression coefficients estimated by applying the conditional likelihood analysis, based on equation (5.28) with $r(\mathbf{x}; \boldsymbol{\beta}) = \exp(\mathbf{x}\boldsymbol{\beta})$, to case-control

samples drawn from the 91 risk sets depicted in Tables 5.3 and 5.4. Twenty controls were sampled from each risk set, except that at age 84 (period 1955–1959) all 17 available controls were used. Subsamples of ten and five were then drawn from the 20. Thus, the errors in the estimated coefficients resulting from the post-hoc sampling are not statistically independent. Comparison of the case-control results with those of the full partial likelihood or parametric analyses shows that the standard errors of the estimated coefficients, especially for heavy arsenic exposure, increase sharply as m (the number of controls) decreases. This reflects the loss in information as fewer members of each risk set are utilized in the analysis. Twenty controls per case seems none too large a number if one wants estimates that are reasonably close to those obtained from the full partial likelihood analysis.

Table 5.7 presents the results of a similar set of case-control analyses, including data for the additional follow-up through 1977, for comparison with the partial likelihood results in Table 5.5. Sets of five, ten and 20 controls were drawn from each of 167 risk sets. The number of data records that were analysed thus approached 3600 when using the maximum number (20) of controls. This limited somewhat the number of exposure variables that could be accommodated, interfered with the interactive nature of the analysis, and thus reduced the advantages of the methodology. The increase in the estimated standard errors as one goes from the full partial likelihood analysis (Table 5.5) to $m = 5$ controls is in the range of 23% to 32% for the regression variables in the

Table 5.7 Regression coefficients and standard errors from case-control analyses of the Montana cohort: 1938–1977

Regression variable	Number of controls (m)			Proportionate mortality (other deaths as controls)
	$m = 20$	$m = 10$	$m = 5$	
All covariables binary (0/1)				
Employed before 1925	0.410 ± 0.163	0.527 ± 0.172	0.349 ± 0.189	0.368 ± 0.169
Foreign-born	0.539 ± 0.168	0.590 ± 0.183	0.452 ± 0.205	0.564 ± 0.181
Moderate arsenic				
1–4 years	0.639 ± 0.181	0.585 ± 0.193	0.594 ± 0.216	0.762 ± 0.198
5–14 years	0.211 ± 0.258	0.192 ± 0.271	0.187 ± 0.292	0.525 ± 0.275
15+ years	0.611 ± 0.227	0.495 ± 0.243	0.585 ± 0.269	0.583 ± 0.249
Heavy arsenic				
1–4 years	0.411 ± 0.337	0.552 ± 0.370	0.482 ± 0.405	0.065 ± 0.353
5+ years	1.228 ± 0.262	1.193 ± 0.288	1.303 ± 0.346	0.867 ± 0.280
−2 × log-likelihood	1432.82	1092.61	794.79	1076.37
Continuous arsenic exposure variables				
Employed before 1925	0.378 ± 0.164	0.502 ± 0.172	0.379 ± 0.188	0.359 ± 0.168
Foreign-born	0.554 ± 0.167	0.589 ± 0.182	0.430 ± 0.202	0.588 ± 0.180
Years moderate arsenic (×10)	0.159 ± 0.075	0.131 ± 0.081	0.140 ± 0.089	0.169 ± 0.082
Years heavy arsenic (×10)	0.538 ± 0.189	0.599 ± 0.256	0.489 ± 0.275	0.417 ± 0.138
−2 × log-likelihood	1448.78	1104.40	809.43	1090.97

second part of the table. The percentage increases in Table 5.6 were larger (25–77%). Theoretical calculations (see §7.6) suggest that the largest increases in standard error should occur with exposures that are relatively infrequent and that have large relative risks. This effect is seen in Table 5.6 but is not so obvious with the updated analysis in Table 5.7.

There is reasonably good agreement between the coefficients of the continuous arsenic exposure variables shown in Table 5.6 and those shown in the second part of Table 5.5, in spite of the fact that the number of respiratory cancer deaths used in the latter analysis was nearly twice that used in the former. However, the relative risk estimated for foreign birth has declined considerably from the earlier analysis. This is due to confounding with date of hire, which is not considered in Table 5.6.

(*f*) *Proportional mortality analyses*

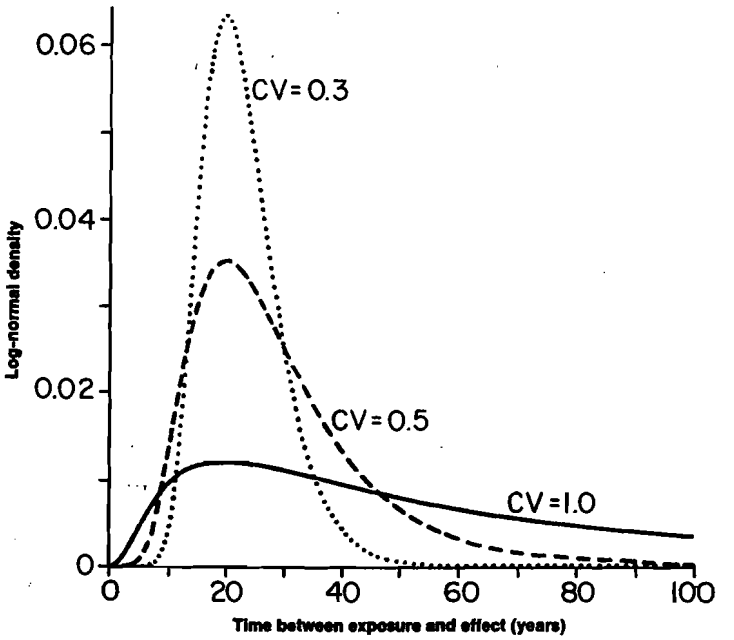
The final columns of Tables 5.6 and 5.7 present results of parallel case-control analyses for the 1938–1963 and 1938–1977 data, respectively, in which the controls consist of all deaths from causes other than respiratory cancer in the 91 or 167 risk sets. These are reasonably comparable with the results of the other case-control analyses. However, the coefficients associated with heavy arsenic exposure generally appear to be smaller, which suggests that heavy arsenic exposure may have adverse effects on mortality from causes other than lung cancer.

(*g*) *Estimating the 'latent interval'*

One of the ways of constructing cumulative exposure functions from a time record of exposure levels is as a time-weighted average (see §5.1). This means selecting the weight function $w(u)$ in equation (5.1) to be a probability density. Several authors have proposed that the log-normal distribution has an intuitively reasonable shape in this context. They assume that there is a random interval of time T between each exposure increment and its effect on the probability of cancer development, and that $\log T$ has a normal distribution with mean μ and σ^2. The corresponding distribution of T has a mode at $\exp(\mu - \sigma)^2$, and its coefficient of variation is $\{\exp(\sigma^2) - 1\}^{1/2}$. Figure 5.7 graphs log-normal density functions with modes at 20 years and various coefficients of variation.

While this manner of constructing exposure functions has a strong intuitive rationale, it is not suggested by any particular biological theory of carcinogenesis, and its use in cancer epidemiology could well be questioned. Nevertheless, largely out of curiosity, we fit a number of models analogous to those shown in the second part of Table 5.7 but in which the cumulative exposure variables were calculated as time-weighted average exposures with log-normal densities. Table 5.8 presents the results. Comparing the goodness-of-fit measures and considering the curves in Figure 5.7, it is clear that strikingly different densities give very similar fits and that precise estimation of the 'latent interval' is simply not possible with this model and these data. The best fit is obtained with a rather peaked distribution (coefficient of variation = 0.1) and a mode at 20 years, but the interpretation of this result is unclear for the reasons already mentioned.

Fig. 5.7 Density functions for the log-normal distribution with mode at 20 years and various coefficients of variation (CV)



This approach is not, of course, limited to the log-normal distribution. Parameters in the other weight functions considered following equation (5.1) could also be varied, to see which gave the best fit.

(*h*) *SMR by years since first employed: case-control approach*

We now return to the analyses depicted in Figures 5.3–5.6, in which we studied the evolution in the respiratory cancer SMR as a function of years since initial employment. The object is to determine empirically how well we can reproduce these results, which required lengthy calculations involving the entire cohort data set, from the samples of the cases in each of the 57 risk sets plus five or 20 controls drawn from the noncases. The illustrative analyses are restricted to estimation of the SMR without covariate adjustment, since this curve had a more distinctive shape than the adjusted curve, even if it was misleading. The results shown are averages of those obtained with 25 separate samplings of five controls per risk set and 15 separate samplings of 20 controls. Elsewhere in this section we have considered results obtained from only a single sampling (as would be done in practice).

Table 5.8 Regression coefficients and standard errors for a series of log-normally time-weighted average exposure models fitted to the Montana cohort data; case-control ($m = 20$) analysis

Regression variable	Coefficient of variation			
	0.5	0.1	0.05	0.0[a]
A. Mode = 15 years				
Foreign-born	0.53 ± 0.16	0.54 ± 0.16	0.54 ± 0.16	0.53 ± 0.16
Moderate arsenic[b]	0.72 ± 0.28	0.52 ± 0.22	0.52 ± 0.22	0.54 ± 0.21
Heavy arsenic[b]	2.12 ± 0.45	1.51 ± 0.35	1.44 ± 0.34	1.42 ± 0.34
−2 × log-likelihood	1519.68	1523.36	1524.04	1523.53
B. Mode = 20 years				
Foreign-born	0.52 ± 0.16	0.53 ± 0.16	0.53 ± 0.16	0.52 ± 0.16
Moderate arsenic	0.88 ± 0.32	0.67 ± 0.24	0.67 ± 0.23	0.70 ± 0.22
Heavy arsenic	2.40 ± 0.53	1.83 ± 0.36	1.75 ± 0.35	1.67 ± 0.35
−2 × log-likelihood	1519.44	1515.44	1515.99	1516.43
C. Mode = 25 years				
Foreign-born	0.52 ± 0.16	0.53 ± 0.16	0.53 ± 0.16	0.52 ± 0.16
Moderate arsenic	1.05 ± 0.37	0.70 ± 0.25	0.68 ± 0.25	0.64 ± 0.23
Heavy arsenic	2.88 ± 0.66	1.86 ± 0.41	1.75 ± 0.38	1.61 ± 0.37
−2 × log-likelihood	1521.32	1519.68	1520.13	1521.26

[a] Exposure effect concentrated on a one-year period 15, 20 or 25 years later
[b] Lagged two years

Figure 5.8A contrasts the curve obtained using all the available data (also shown in Figure 5.6) with the average curves obtained by applying the Taylor series (5.30) and jackknife (5.31) estimates to case-control samples with five controls per risk set. The bias is clearly unacceptable, the Taylor series estimate overestimating the SMR and the jackknife underestimating it for the first 20–30 years. A much more satisfactory result is obtained by using 20 controls per risk set (Fig. 5.8B) or by pooling the five risk sets containing five controls each for which the t_i are within 2.5 years of the target value (Fig. 5.8C). The latter procedures both provide a reasonably faithful reproduction of the original result.

5.6 Continuous variable analysis of nasal sinus cancer deaths among Welsh nickel refiners

We continue our analyses of cohort data from the Welsh nickel refiners study in order to illustrate some further features of continuous variable modelling. These data have already been considered in Example 4.1 and §4.10.

Table 5.9 presents observed and expected numbers of nasal sinus cancer deaths according to the four risk variables of primary interest: age at first employment, year of employment, exposure index and time since first employment. Many of the essential features of the data are already evident from these simple descriptive statistics. The

Fig. 5.8 Smoothed estimates of the standardized mortality ratio (SMR) for respiratory cancer for Montana smelter workers estimated from 15 case-control samples. (A) Five controls per risk set, no pooling; (B) 20 controls per risk set, no pooling; (C) five controls per risk set with pooling of five neighbouring risk sets. ——, all controls; – – –, jackknife; —·—, Taylor series

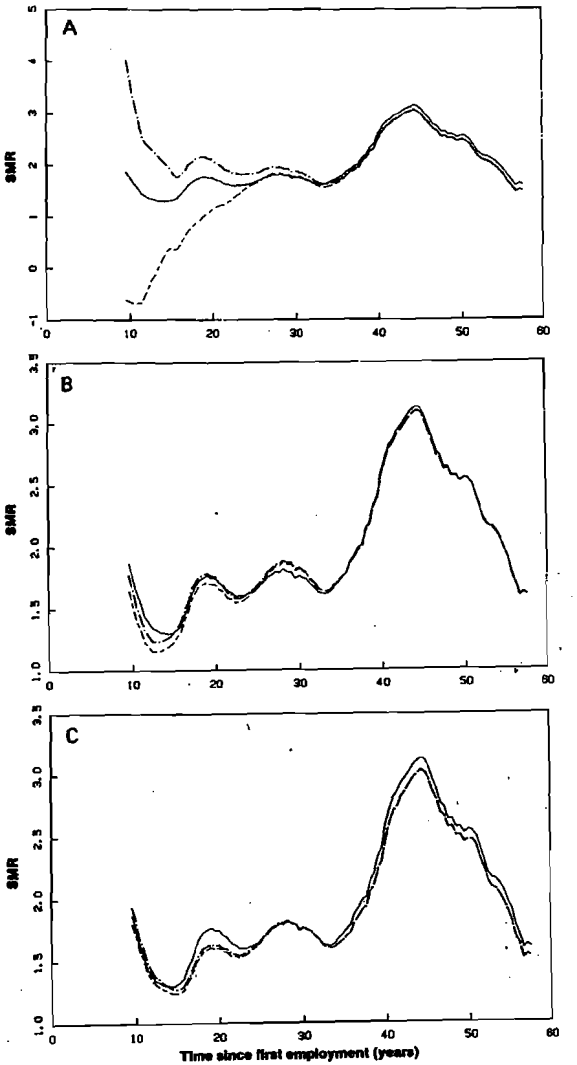


Table 5.9 Summary data on deaths from nasal sinus cancer among Welsh nickel refiners[a]

Variable	Category	Person–years	Nasal sinus cancer		
			Observed	Expected	Rate[b]
Age at first employment (years)	15–19	3 089.2	2	0.029	0.6
	20–24	4 773.9	11	0.057	2.3
	25–29	4 186.5	18	0.065	4.3
	30–34	1 816.3	11	0.031	6.1
	35–39	986.1	12	0.018	12.2
	40–44	233.9	1	0.006	4.3
	45+	144.9	1	0.003	6.9
Year of first employment	1900–1904	277.4	2	0.007	7.2
	1905–1909	1 673.6	9	0.033	5.4
	1910–1914	2 904.5	26	0.045	9.0
	1915–1919	2 294.0	9	0.030	3.9
	1920–1924	8 081.3	10	0.095	1.2
Exposure category (years)	0.0	7 738.8	10	0.102	1.3
	0.5–4.0	4 905.1	17	0.065	3.5
	4.5–8.0	1 716.9	12	0.027	7.0
	8.5–12.0	601.1	10	0.010	16.6
	12.5+	268.9	7	0.004	26.0
Time since first employed (years)	15–19	2 586.1	1	0.011	0.4
	20–24	2 194.3	3	0.016	1.4
	25–29	2 583.2	16	0.028	6.2
	30–34	2 379.3	10	0.033	4.2
	35–39	1 950.0	7	0.032	3.6
	40–44	1 426.2	5	0.028	3.5
	45–49	1 035.2	8	0.027	7.7
	50–54	652.9	5	0.020	7.7
	55+	423.5	1	0.014	2.4
Totals		15 230.8	56	0.210	3.7

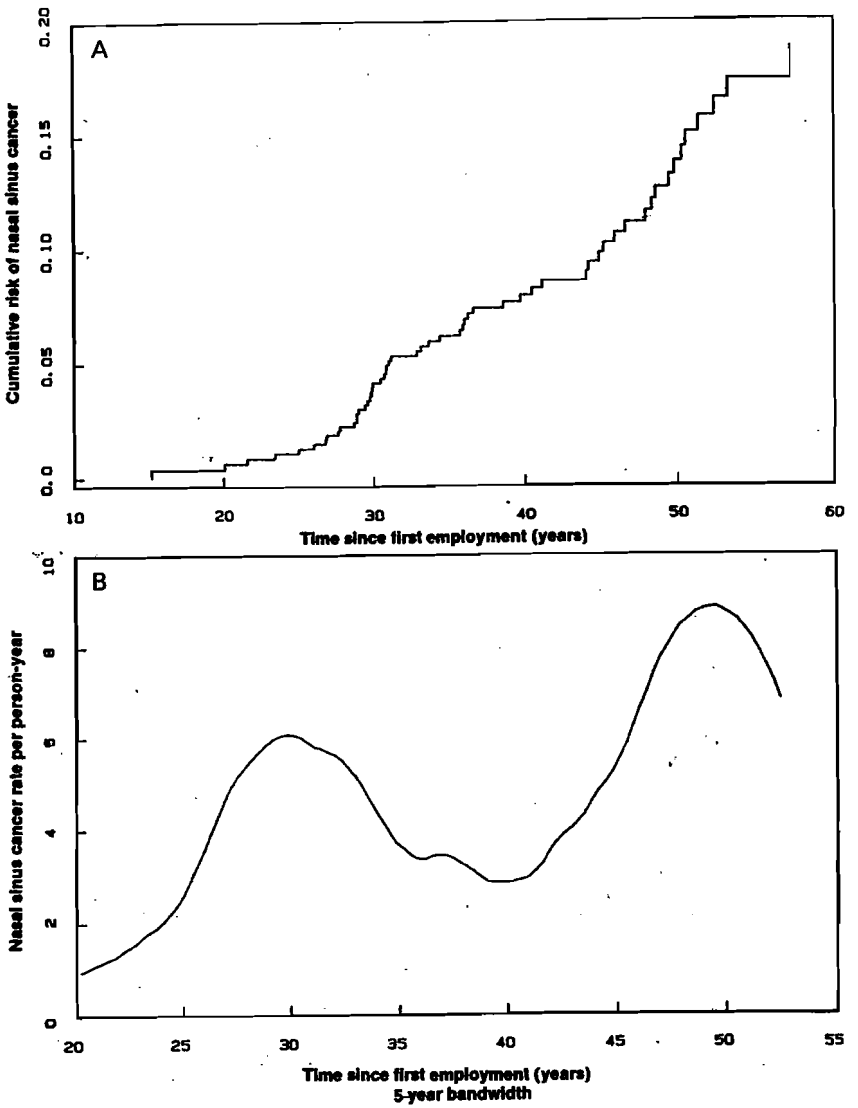
[a] Determined from data shown in Appendix VIII. There are slight differences between Tables 4.23 and 5.9 in the totals of expected numbers of deaths due to the use of slightly different data.
[b] Nasal sinus cancer death rate per 1000 person–years of observation

nasal sinus cancer rates increase dramatically with duration of 'exposure': they seem to have one peak about 25–29 years from date of hire and another at about 50 years. However, both exposure and time since employment are correlated with age and year of employment. A major goal of our analysis will be to try to separate the effects of each of the explanatory variables using an appropriate regression model. Results for nasal sinus cancer are analysed without reference to standard rates since the 'background' is so inconsequential.

(a) *Analysis of nasal sinus cancer risk by time since first employment*

Figure 5.9A graphs the cumulative nasal sinus cancer death rate for the entire cohort as a function of time since initial employment (equation 5.16). We estimate the

Fig. 5.9 Death rate from nasal sinus cancer by years since initial employment, Welsh nickel refiners. (A) Cumulative rate; (B) smoothed instantaneous rate



cumulative 'lifetime' risk (to age 85) to be about 15–20%, a striking figure when one considers how rare the disease is in the general population. The smoothed estimate of the annual death rates shown in part B of the figure, obtained from equation (5.18) with a five-year bandwidth, confirms the possibility of a bimodal pattern that was already evident in the grouped data of Table 5.9.

Our initial analysis of these data used the proportional hazards model with log-linear relative risk function (equation 5.3) and considered t = 'time since first employment' (TFE) as the basic time variable. We define a number of indicator variables to identify levels of the factors age at first employment (AFE), year of first employment (YFE) and exposure (EXP). Recall that AFE, YFE and TFE were investigated jointly in the grouped data analysis of Example 4.1.

Since ages were recorded to two-decimal accuracy, and we retained this level of detail in the analysis, each of the 56 cases of nasal sinus cancer occurred at a unique time since first employment and generated a separate risk set. The first case occurred at 15.23 years from initial hire, at which time there were 284 individuals under observation. Risk-set sizes increased gradually to a maximum of 531 men at risk at 28.72 years since date of hire and then declined. The smallest risk set, with 73 subjects, was at 57.48 years since date of hire, the maximum number of years at which a case was observed.

Table 5.10 summarizes the results of fitting the model with categorical regression variables by partial likelihood. Each of the factors AFE, YFE and EXP is seen to have a strong, independent effect on risk. The rise in relative risk with age at first employment is a particularly striking and unusual observation (Peto, J. *et al.*, 1984). While an increase in risk with AFE is evident in the summary data of Table 5.9, its magnitude is obscured by the fact that those hired at later ages generally did not survive to the point 45–50 years from date of employment at which the nasal sinus cancer rates are highest. Once this confounding is accounted for in the regression analyses, the role of AFE appears to be even more dramatic.

The baseline cumulative death rate is shown graphically in Figure 5.10A; a smoothed estimate of the instantaneous death rate, using a five-year bandwidth, appears in part B of the figure, and for a ten-year bandwidth in part C. Because of the coding of the covariables, this baseline risk is estimated for a (fictitious) subject who was under 20 years at hire, first worked before 1910 and was never assigned to high-risk categories. The estimated cumulative lifetime risk for this category does not exceed 1%, whereas for the cohort as a whole it approaches 15–20%. Furthermore, the peak in the nasal sinus cancer death rate at 30 years past employment that was suggested by the crude analysis (Fig. 5.9B) essentially dissappears when adjustment is made for the covariable effects.

The second part of Table 5.10 reports the fit of a model with continuous rather than discrete covariables. The definitions of the covariables used in this fit were determined after considering the results in the first part of the table and after conducting some exploratory analyses using the case-control technique (see below). Comparing the maximized partial likelihoods obtained from the grouped and continuous analyses, we conclude that the fit with four continuous covariables is almost as good as that with the larger number of binary variables that identified categories of risk.

Table 5.10 Results of fitting the multiplicative model by maximum likelihood to data on nasal sinus cancer deaths; 'time' = years since first employed

Variable[a]	Level	Parameter estimate ± standard error	p value	Relative risk
All covariables discrete				
AFE (years)	15–19			1.0
	20–27.5	1.48 ± 0.75	0.05	4.4
	27.5–35	2.21 ± 0.76	0.004	9.1
	35+	3.64 ± 0.79	0.00004	38.0
YFE	1900–1909			1.0
	1910–1914	1.03 ± 0.38	0.007	2.8
	1915–1919	1.11 ± 0.51	0.03	3.0
	1920–1924	0.01 ± 0.53	0.98	1.0
EXP (years)	0			1.0
	0.5–4.0	0.88 ± 0.40	0.03	2.4
	4.5–8.0	1.19 ± 0.47	0.01	3.3
	8.5–12.0	2.30 ± 0.52	0.001	10.0
	12.5+	2.84 ± 0.57	0.0001	17.2
		−2 × log-likelihood = 561.2		
All covariables continuous				
log(AFE-10)		2.22 ± 0.44	<0.00001	
(YFE-1915)/10		−0.09 ± 0.32	0.76	
(YFE-1915)²/100		−1.26 ± 0.51	0.01	
log(EXP + 1)		0.77 ± 0.17	0.00001	
		−2 × log-likelihood = 568.9		

[a] AFE, age at first employment; YFE, year of first employment; EXP, duration of 'exposure' in designated job categories

One goal of constructing appropriate continuous covariables was to lay the foundation for assessing the goodness-of-fit of the multiplicative model by incorporating cross-product or interaction terms in the regression equation. Such analyses are more sensitive if the interactions can be expressed in a quantitative rather than a qualitative manner so that the chi-square statistics for testing their significance have at most a few degrees of freedom (see Volume 1, §6.6 and 6.7). For reasons of economy and convenience, however, these explorations for interaction effects were restricted to the case-control analyses reported below.

(*b*) *Analysis of nasal sinus cancer risk by attained age*

We did not choose attained age as the basic time variable in our initial partial likelihood analysis of the nasal sinus cancer deaths. Since it is obvious that all or nearly all such cases were caused by the specific nickel exposure rather than by general environmental exposures, the usual reasons for regarding age as the key explanatory variable were absent. Most persons concerned with the analysis of these data have considered duration of time since onset of exposure to the causal agent to be the most relevant time scale.

Fig. 5.10 Baseline death rate from nasal sinus cancer by years since initial employment for Welsh nickel refiners, estimated by the multiplicative model. (A) Cumulative rate; (B) smoothed instantaneous rate (five-year bandwidth); (C) smoothed instantaneous rate (ten-year bandwidth)

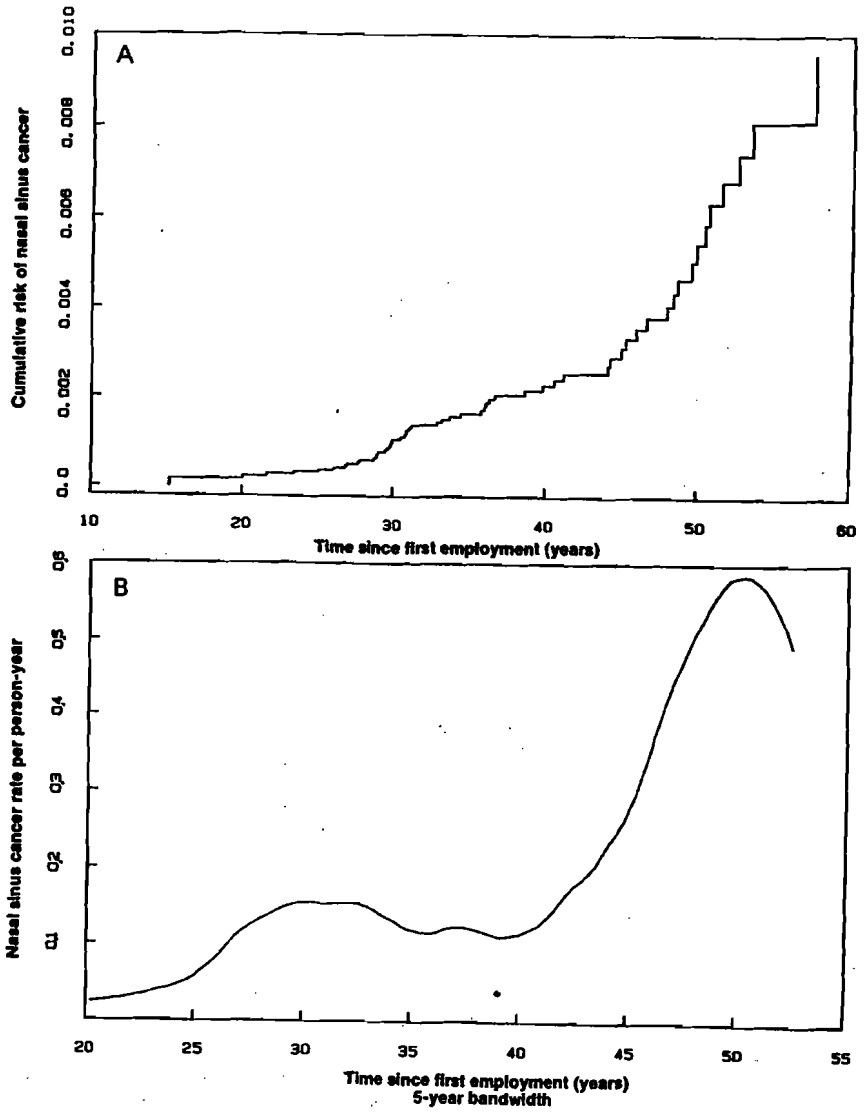
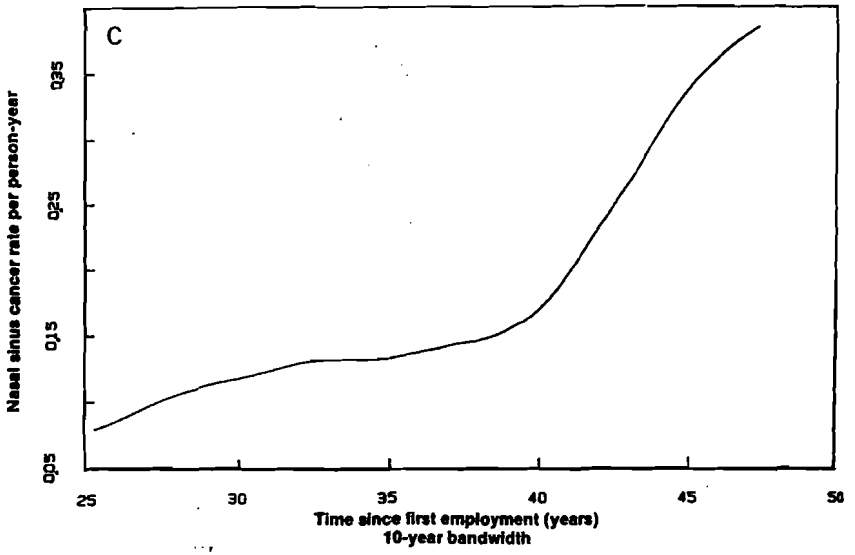


Fig. 5.10 (*contd*)



One might ask whether attained age should be included as an additional variable in the analysis to see whether it carries some explanatory value after accounting for age at onset and time since first exposure. However, since attained age is the sum of these latter two variables, it is clear that such an analysis cannot separate the (linear) effects of the three factors. The basic problem is the same as that which occurs also with age-period-cohort analyses.

Nevertheless, largely out of curiosity, we did conduct an alternative partial likelihood analysis with attained age *replacing* time since first employment as the basic time variable. Table 5.11 reports the regression coefficients for the discrete levels of the factors AFE, YFE and EXP obtained with this approach, and Figure 5.11 shows the smoothed estimate of the baseline death rate as a function of age. The relative risks associated with YFE and EXP depend little on whether the baseline risk is expressed as a function of age or of time since onset of exposure. The increase in relative risk with AFE, however, is substantially less when age is used as the basic time variable. Correspondingly, the baseline risk increases more smoothly and sharply as a function of age than as a function of time since onset of exposure. (Compare Figures 5.10B and 5.11.)

(*c*) *Nasal sinus cancer deaths: sampling from the risk set*

In order to reduce the volume of data so as to explore different ways of constructing continuous regression variables and to search for significant interactions, we carried out

Table 5.11 Results of fitting the multiplicative model by maximum partial likelihood to data on nasal sinus cancer deaths; 'time' = age

Variable[a]	Level	Parameter estimate ± standard error	*p* value	Relative risk
AFE (years)	15–19			1.0
	20–27.5	1.03 ± 0.75	0.17	2.8
	27.5–35	1.30 ± 0.75	0.08	3.7
	35+	2.08 ± 0.77	0.007	8.0
YFE	1900–1909			1.0
	1910–1914	0.93 ± 0.38	0.01	2.5
	1915–1919	0.93 ± 0.51	0.07	2.5
	1920–1924	−0.12 ± 0.52	0.82	0.9
EXP (years)	0			1.0
	0.5–4.0	0.82 ± 0.40	0.04	2.3
	4.5–8.0	1.10 ± 0.47	0.02	3.0
	8.5–12.0	2.24 ± 0.51	0.0001	9.4
	12.5+	2.77 ± 0.57	0.00001	16.1
		−2 × log-likelihood = 573.36		

[a] See legend to Table 5.10

Fig. 5.11 Adjusted nasal sinus cancer rates by age, smoothed using five- (——) and ten-year (– – –) bandwidths

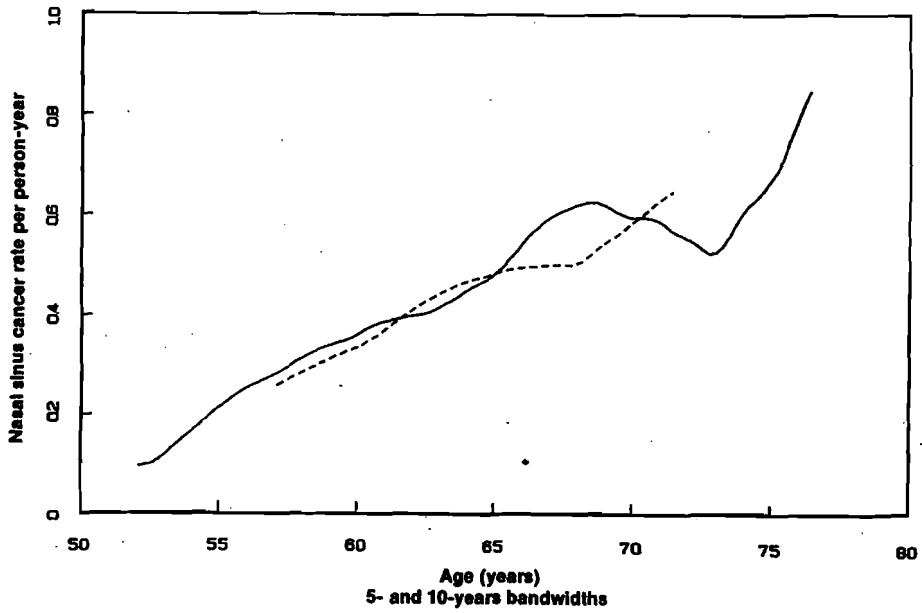


Table 5.12 Results of fitting the multiplicative models by conditional maximum likelihood to matched sets of a nasal sinus cancer case and 20 controls; 'time' = years since first employed

Variable[a]	Level	Parameter estimate ± standard error	*p* value	Relative risk
All covariables discrete				
AFE (years)	15–19			1.0
	20–27.5	1.52 ± 0.76	0.048	4.5
	27.5–35	2.12 ± 0.78	0.006	8.2
	35+	3.60 ± 0.83	<0.001	36.7
YFE	1900–1909			1.0
	1910–1914	0.74 ± 0.40	0.064	2.1
	1915–1919	0.85 ± 0.53	0.127	2.3
	1920–1924	−0.30 ± 0.53	0.571	0.7
EXP (years)	0			1.0
	0.5–4.0	0.83 ± 0.42	0.049	2.3
	4.5–8.0	0.93 ± 0.48	0.049	2.5
	8.5–12.0	2.45 ± 0.56	<0.001	11.6
	12.5+	2.56 ± 0.63	<0.001	13.0
		Deviance = 259.80		
All covariables continuous				
log(AFE-10)		2.09 ± 0.46	<0.001	
(YFE-1915)/10		−0.23 ± 0.32	0.438	
$(YFE-1915)^2/100$		−1.01 ± 0.53	0.057	
log(EXP + 1)		0.72 ± 0.18	<0.001	
		Deviance = 267.67		

[a] See legend to Table 5.10

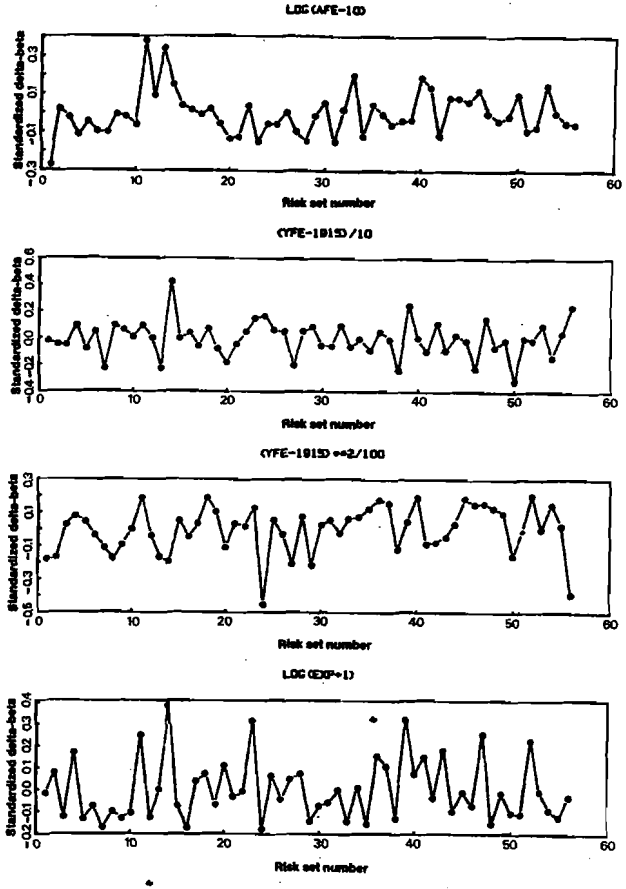
Table 5.13 Deviances for various interaction terms when fitting the multiplicative model to matched sets of a nasal sinus cancer case and 20 controls; 'time' = years since first employed

Interaction variables included in equation[a]	Deviance
None	267.67
log (AFE-10) × (YFE-1915)/10 + log (AFE-10) × $(YFE-1915)^2/10$	267.22
log (AFE-10) × log (EXP + 1)	267.51
(YFE-1915)/10 × log (EXP + 1) + $(YFE-1915)^2/100$ × log (EXP + 1)	267.04
log (AFE-10) × log (AFE-10)	267.62
log (EXP + 1) × log (EXP + 1)	266.78
log (AFE-10) × TFE	266.92
(YFE-1915)/10 × TFE + $(YFE-1915)^2/100$ × TFE	265.41
log (EXP + 1) × TFE	267.66

[a] In addition to four continuous variables shown in the second part of Table 5.12; see legend to Table 5.10

the risk-set sampling procedure, selecting 20 controls from each of the 56 risk sets. This yielded a data file containing $21 \times 56 = 1176$ records that could be analysed with relative ease. Table 5.12 shows the results of fitting the same models to the case-control data as had been fitted earlier to the entire data set (Table 5.11). While there is reasonably good agreement, the relative risks associated with the highest exposure category and employment in the 1910–1919 period are underestimated with the case-control data. Just as we found for the full analysis, however, a summary of the data in terms of the four continuous variables is quite adequate in comparison with a summary in terms of the corresponding discrete factors.

Fig. 5.12 Deletion diagnostics for the model shown in the second part of Table 5.12; approximate effect on the standardized regression coefficients from deletion of individual cases. AFE, age at first employment; YFE, year of first employment; EXP, duration of 'exposure' in designated job categories



The question of possible interactions between the continuous variables which, if present, would tend to invalidate the results obtained with the simple multiplicative model is examined in Table 5.13. For no risk variable was there any indication of a (linear) dependence of its multiplicative effect on values of another risk variable or on time since first employment, nor was there strong evidence of curvature in the dependence of log relative risk on log (AFE − 10) or log (EXP + 1). Had there been, it would suggest that some other transformation of these variables be used instead.

One last check on the adequacy of the fitted model was to examine the approximate change in the regression coefficients estimated for each of the four continuous variables that would accompany the deletion of any one of the 56 cases from the analysis. Since each risk set contained a single case, deletion of a case has the same effect as deleting the entire risk set for these data. Results obtained using the procedure of Storer and Crowley (1985) are shown in Figure 5.12. The risk sets are numbered according to time since first employment so that number 1 corresponds to the case diagnosed at 15.23 years and number 56 to the case diagnosed at 57.48 years. For none of the four variables does deletion of a risk set change the estimated value of the β regression coefficient by more than half its standard error. The linear and square terms in YFE are correlated, so that the deletion of certain risk sets (e.g., numbers 14, 24, 56) causes the coefficient of (YFE-1915)/10 to increase and that for $(\text{YFE-1915})^2/100$ to decrease, and *vice versa*.