4. FITTING MODELS TO GROUPED DATA

CHAPTER 4

FITTING MODELS TO GROUPED DATA

A major goal of the statistical procedures considered in the preceding two chapters was to condense the information in a large set of incidence or mortality rates into a few summary measures so as to estimate the effects that a risk factor has on the rates. A secondary goal was to evaluate the statistical significance of the effect estimates at different levels of exposure in order to rule out the possibility that the observed differences in rates were due simply to the play of chance. Some attention was devoted also to determining whether the effect measures used (relative risks) were reasonable summary measures in the sense of remaining relatively constant from one age stratum to the next, or whether, instead, it was necessary to describe how the effect was modified by age or other variables used for stratification.

Role of statistical modelling

Estimation of risk factor effects and tests of hypotheses about them are also the goals of statistical modelling. The statistician constructs a probability model that explicitly recognizes the role of chance mechanisms in producing some of the variation in the rates. Observed rates are regarded as just one of many possible realizations of an underlying random process. Parameters in the model describe the systematic effects of the exposures of interest, and estimates of those parameters, obtained during the process of fitting the model to the data, serve as summary statistics analogous to the SMR or Mantel–Haenszel estimates of relative risk. Evaluation of dose-response trends is conducted in terms of tests for the significance of regression coefficients for variables representing quantitative levels of exposure. Additional parameters may be incorporated in order to model variations of the exposure effects with age, calendar year or other stratum variables.

Statistical modelling has several advantages over standardization and related techniques. It facilitates consideration of the simultaneous effects of several different exposure variables on risk. Applied to the study of nasal sinus and lung cancers in Welsh nickel workers, for example, the effects of period of employment, age at employment and years since employment may be estimated in a single model equation (see §4.3) rather than in separate stratified analyses (Tables 3.12 and 3.13). If quantitative variables are available that specify the timing and degree of exposure, then a more economical description of the data often may be given in terms of dose-time-response relationships rather than by making separate estimates of risk for

each exposure category. Such quantitative expression of the results facilitates the interpolation of risk estimates for intermediate levels of exposure. It is essential for extrapolation beyond the range of the available data, although this is usually a hazardous undertaking. Examination of the goodness-of-fit of the model to the observed rates alerts the investigator to situations in which the simple model description is inadequate or in which important features of the data are being overlooked. Estimates of relative risk obtained by model fitting generally have greater numerical stability than those computed from standardized rates.

There are, of course, some apparent drawbacks to model fitting that need to be considered along with the advantages. Perhaps the greatest problem lies in the parametric specification of the model. While explicit theories about the nature of the disease process are sometimes available to suggest models with a particular mathematical form (see Chapter 6), more often the models used in statistical data analysis are selected on the basis of their flexibility and because the associated fitting procedures are well understood and convenient. Alternative models may have quite different epidemiological interpretations. Examining the relative goodness-of-fit of two distinct model structures enables one to judge whether the evidence favours one interpretation over another, or whether they are both more or less equally in agreement with the observed facts. Unfortunately, epidemiological data are rarely extensive enough to be used to discriminate clearly between closely related models, and some uncertainty and arbitrariness in the process of model selection is to be anticipated. Nevertheless the very act of thinking about the possible biological mechanisms that could have produced the observations under study can be beneficial. Consideration of possible model structures is not strictly necessary when applying the elementary techniques, but even these implicitly assume some regularity in the basic data and, as we have seen, may yield misleading answers if it is absent.

Scope of Chapter 4

This chapter develops methods for the analysis of grouped cohort data that are based on maximum likelihood estimation in Poisson models for the underlying disease rates. Additive and multiplicative models are introduced in §4.1 as a means of summarizing the basic structure in a two-dimensional table of rates. It is again shown that the ratio of two CMFs appropriately summarizes age-specific rate ratios under the multiplicative model, but that the ratio of two SMRs does not unless additional assumptions are met. The basic process of model fitting is illustrated by an analysis of Icelandic breast cancer rates classified by age and birth cohort.

Section 4.2 contains more technical material that justifies the use of the Poisson model as the basis for maximum likelihood analysis of grouped cohort data. It may be omitted on a first reading.

Methods of fitting multiplicative models to grouped cohort data consisting of a multidimensional cross-classification of cases (or deaths) and person-years denominators are developed in §4.3. The computer program GLIM is shown to offer particularly convenient features for fitting Poisson regression models. Quantities available from the GLIM fits are easily converted into 'deletion diagnostics' that aid in

assessing the stability of the fitted model under perturbations of the basic data. These techniques are by no means limited to the analysis of relative risk: §4.4 shows that GLIM may be used also to fit a class of generalized linear models that range from additive to multiplicative. Methods for selecting the model equation that best describes the structure in the data are illustrated by application to a rather simple problem involving coronary deaths among smoking and nonsmoking British doctors.

The Montana smelter workers data from Appendix V are reanalysed in §§4.5, 4.6 and 4.7 in order to demonstrate the close connection between multiplicative models and the elementary techniques of standardization and Mantel–Haenszel estimation introduced in §§3.4, 3.6 and 3.7. Section 4.5 considers internal estimation of background rates from study data, whereas §4.6 develops analogous models that incorporate external standard rates. Proportional mortality analyses based on fitting of logistic regression models to case-'control' data, both with and without reference to external standard proportions, are developed in §4.7.

More comprehensive analyses of the Montana data, using original records not published here, appear in §§4.8 and 5.5. Some additional models that do not fall strictly under the rubric of the generalized linear model are considered in the last two sections of the chapter. Foremost among these is the additive relative risk model whereby different exposures act multiplicatively on the background rates, but combine additively in determining the relative risk. This is illustrated in §4.9 by application to data on lung cancer deaths among British doctors. GLIM macros are presented for fitting a general class of relative risk models which includes both the additive and multiplicative as special cases. In §4.10, grouped data from the Welsh nickel refiners study are used to illustrate the fitting of a model in which the excess risk of lung cancer (over background based on national rates) is expressed as a mutliplicative combination of exposure effects. These results are contrasted with those of a more conventional multivariate analysis of the SMR under the multiplicative model.

Some familiarity with the principles of likelihood inference and linear models is assumed. Readers without such background are referred to §§6.1 and 6.2 of Volume 1, and the references contained therein, for an appropriate introduction.

4.1 Additive and multiplicative models for rates

Most of the essential concepts involved in statistical modelling can be introduced by considering the simple example of a two-dimensional table of rates. The data layout (Table 3.4) consists of a table with J rows ($j = 1, \ldots, J$) and K columns ($k = 1, \ldots, K$). Within the cell formed by the intersection of the jth row and kth column, one records the number of incident cases or deaths d_{jk} and the person-years denominators n_{jk}. For concreteness, we may think of j as indexing J age intervals and k as representing one of K exposure categories.

The observed rate in the (j, k)th cell may be written $\hat{\lambda}_{jk} = d_{jk}/n_{jk}$. This is considered as an estimate of a true rate λ_{jk} that could be known exactly only if an infinite amount of observation time were available. In order to account for sampling variability, the d_{jk} are regarded as independent Poisson variables with means and variances $E(d_{jk}) = \text{Var}\,(d_{jk}) = \lambda_{jk} n_{jk}$. The denominators n_{jk} are assumed to be fixed. The rationale for this Poisson assumption is discussed in §§4.2 and 5.2.

The goal of the statistical analysis is to uncover the basic structure in the underlying rates λ_{jk}, and, in particular, to try to disentangle the separate effects of age and exposure. This is accomplished by introducing one set of parameters or summary indices which describe the age effects and another set for the exposures. However, such a simple description makes sense only if the age-specific rates display a degree of consistency such that, within defined limits of statistical variation, the relative position of each exposure group remains constant over the J age levels (see Chapter 2, Volume 1). If one exposure group has higher death rates among young persons, but lower rates among the elderly, use of a single summary rate (or the analogous parameter in a statistical model) to represent the exposure effect will obscure the fact that the effect depends on age.

(*a*) *The model equations*

Various possible structures for the rates satisfy the requirement of consistency. In particular, it holds if the effect of exposure at level k is to add a constant amount β_k to the age-specific rates λ_{j1} for individuals in the baseline or nonexposed category ($k = 1$). The model equation is

$$\lambda_{jk} = \alpha_j + \beta_k, \tag{4.1}$$

where $\alpha_j = \lambda_{j1}$ and β_k ($\beta_1 = 0$) are parameters to be estimated from the data.

If additivity does not hold on the original scale of measurement, it may hold for some transformation of the rates. The log transform

$$\log \lambda_{jk} = \alpha_j + \beta_k \tag{4.2}$$

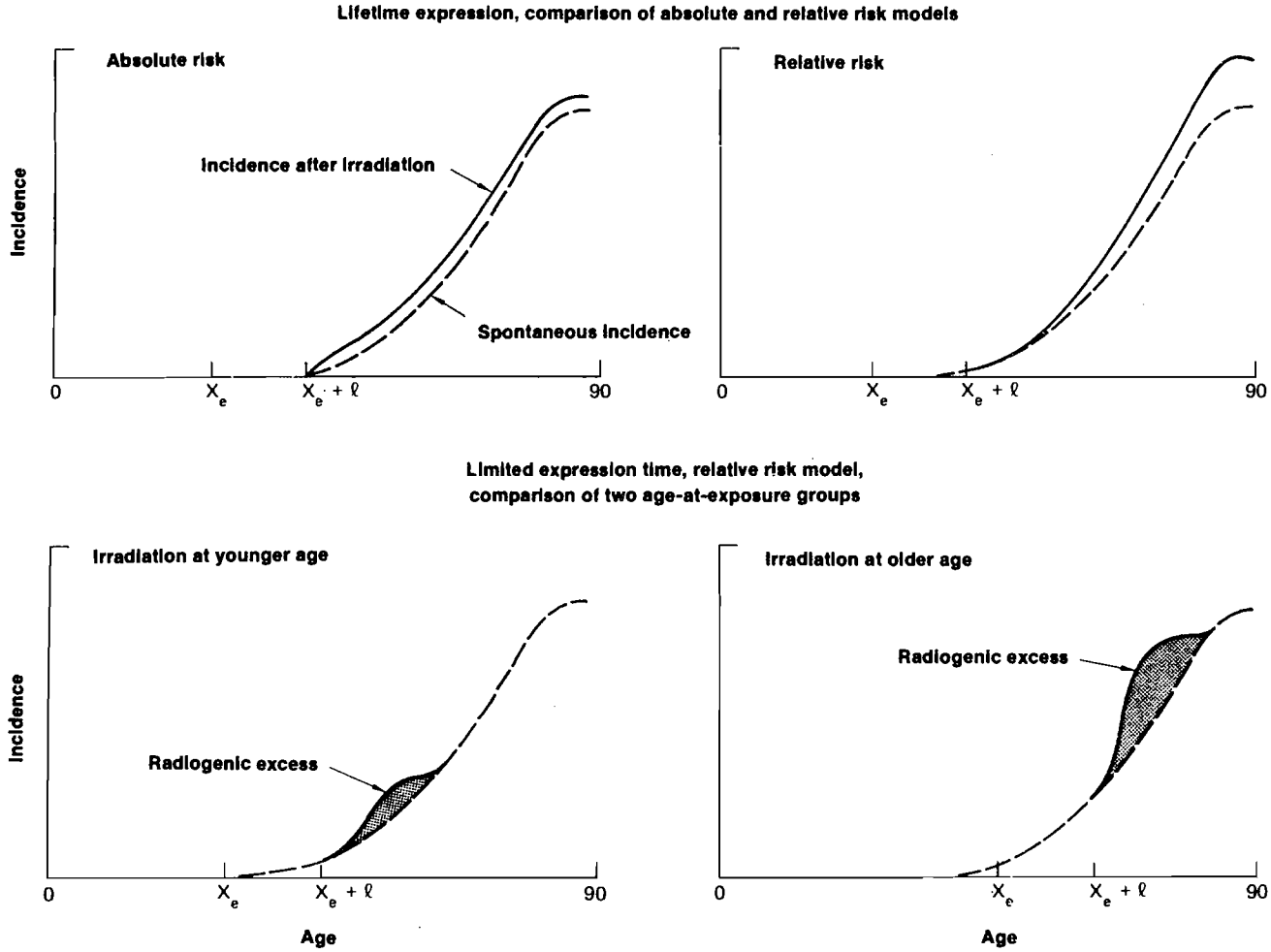
yields the multiplicative model

$$\lambda_{jk} = \theta_j \psi_k,$$

where now $\alpha_j = \log \theta_j = \log \lambda_{j1}$ and $\beta_k = \log \psi_k$. In this case, ψ_k represents the relative risk (rate ratio) of disease for exposure at level k relative to a baseline at level 1 ($\psi_1 = 1$).

The excess (additive) and relative (multiplicative) risk models are the two most commonly used to describe the relationship between the effects of exposure and the effects of age and other 'nuisance' factors that may account for background or spontaneous cases. Both have been used to describe different aspects of radiation carcinogenesis in human populations (Committee on the Biological Effects of Ionizing Radiation, 1980). The upper two panels of Figure 4.1 contrast the age-incidence curves that result from the two models when a given dose of radiation produces a constant effect that persists for life after a latent period. Due to the sharp rise in background incidence with age, relative risk estimates derived from current data generally predict a greater lifetime radiation risk than do estimates of additive effect. The two lower panels of Figure 4.1 illustrate the effect of age at irradiation on risk for a multiplicative model in which the radiation effect itself is concentrated in the period from l_1 to l_2 years after exposure. However, this complication of a limitation of the period of effect is not considered further in this section.

Fig. 4.1 Radiation-induced cancer effect superimposed on spontaneous cancer incidence by age. Illustrations of various possibilities; X_e, age at exposure; l, minimal latent period. From Committee on the Biological Effects of Ionizing Radiation (1980)



Considerable attention has been given in recent years to the problems of discriminating between additive and multiplicative models using epidemiological data (Gardner & Munford, 1980; Thomas, D.C., 1981; Walker & Rothman, 1982; Breslow & Storer, 1985). One possible approach is presented in §4.5. Unless the data are quite extensive and the effect of exposure pronounced, however, random sampling errors may make such discriminations difficult. Furthermore, errors of misclassification of the exposure variable may operate to distort the true relationship (Tzonou *et al.*, 1986). In view of such uncertainties, the choice of model is legitimately based as much on a-priori considerations as it is on goodness-of-fit tests, unless of course these show one or the other model to be markedly superior. As with the report of the Committee on the Biological Effects of Ionizing Radiation, some authors follow the prudent course of examining and presenting their data using several alternative model assumptions.

(*b*) *Biological basis for model selection*

Sections 2.4–2.7 of Volume 1 gave both empirical and logical reasons for the usually greater convenience in cancer epidemiology of measuring the effects of exposures in terms of the relative risk parameters of the multiplicative model rather than the excess risk parameters of the additive model. Incidence and death rates for cancers of epithelial tissue are known to rise rapidly with age, the age-incidence curves approximating a power function with exponent between four and five (Doll, 1971). When plotted on log paper for different exposure or population groups, the age-incidence curves are therefore roughly linear with a common slope but varying intercept (Fig. 2.2). This implies a multiplicative relationship.

If the two dimensions of the table correspond to two different exposure factors, however, then various models for the disease process suggest that their individual effects on the age-specific rates or on the lifetime risks may combine additively, multiplicatively or in some other fashion. Models based on the multistage theory of carcinogenesis lead to approximately additive structures if the two risk factors affect the same stage of the process and to multiplicative structures if two distinct stages are affected (Lee, 1975; Siemiatycki & Thomas, 1981; Hamilton, 1982). A detailed discussion of quantitative theories of carcinogenesis and how they may be used to suggest appropriate dose-time-response relationships involving one or more agents is given in Chapter 6. Under Rothman's (1976) component-sufficient cause paradigm of disease causation, which is perhaps of greater relevance to other areas of epidemiology, 'independent' factors or those which contribute to different disease pathways have effects that combine in a nearly additive fashion, whereas the effects of 'complementary' factors or those that contribute different parts to the same pathway combine in a manner that is close to multiplicative (Koopman, 1982).

(*c*) *Standardization and multiplicative models*

The CMF and the SMR (see Chapter 2) were originally developed from a general, intuitive perspective, in the absence of any formal assumption about the structure that might be present in the underlying age-specific disease rates. Nevertheless, considerable insight into the properties of such statistical measures is gained by investigating their performance under well-defined and plausible models for the basic data. Here, we compare the performance of the CMF and SMR in the multiplicative environment and develop an interesting relationship between the iterative fitting of multiplicative models and the calculation of the indirectly standardized SMR. Similar investigations have been undertaken by Freeman and Holford (1980), Anderson *et al.* (1980) and Hoem (1987).

Suppose, for simplicity, that there are only two exposures categories ($k = 1$ or 2) and denote by $w_j = n_{j0}/N_0$ and $\lambda_j^* = d_{j0}/n_{j0}$ the standard weights and rates that enter into the calculation of the summary measures. According to (4.2) the ratio of age-specific rates for the two categories is equal to ψ_2/ψ_1, or just ψ_2 if $\psi_1 = 1$ as is generally assumed, regardless of the age interval. Thus, the ratio of the two corresponding summary measures should tend towards ψ_2 in large samples if the measures are to reflect accurately the basic regularity in the rates. An easy calculation shows this is indeed

true for direct standardization:

$$\frac{\mathrm{CMF}_2}{\mathrm{CMF}_1} \rightarrow \frac{\sum_{j=1}^{J} w_j \lambda_{j2}}{\sum_{j=1}^{J} w_j \lambda_{j1}} = \frac{\psi_2 \sum_{j=1}^{J} w_j \theta_j}{\psi_1 \sum_{j=1}^{J} w_j \theta_j} = \psi_2.$$

For the ratio of two SMRs, however, we have

$$\frac{\mathrm{SMR}_2}{\mathrm{SMR}_1} \rightarrow \frac{\sum_{j=1}^{J} n_{j2}\lambda_{j2} / \sum_{j=1}^{J} \lambda_j^* n_{j2}}{\sum_{j=1}^{J} n_{j1}\lambda_{j1} / \sum_{j=1}^{J} \lambda_j^* n_{j1}} = \psi_2 \times \frac{\sum_{j=1}^{J} n_{j2}\theta_j / \sum_{j=1}^{J} \lambda_j^* n_{j2}}{\sum_{j=1}^{J} n_{j1}\theta_j / \sum_{j=1}^{J} \lambda_j^* n_{j1}}. \tag{4.3}$$

The second term in this expression generally does not equal 1 unless we also have $\theta_j = \text{const} \times \lambda_j^*$ or else $n_{j2} = \text{const} \times n_{j1}$; that is, unless the age-specific rates for exposure categories 1 and 2 are both proportional to the external standard rates, in addition to being proportional to each other, or else the two age distributions are identical. The bias in the ratio of SMRs can be severe if these conditions are grossly violated, as Table 2.13 makes clear.

The condition of proportionality with the external standard automatically holds for the multiplicative model if one takes for the 'standard' either one of the two sets of age-specific rates that are being compared. If the first exposure group ($k = 1$) is taken as standard for computation of the CMF, and the second group ($k = 2$) as standard for the SMR, then the ratios of CMFs and SMRs are identical (Anderson *et al.*, 1980, Section 7A.4). Using the pool of the two comparison groups as an internal standard, however, generally does not satisfy the proportionality condition, and the ratio of SMRs computed on this basis does not estimate the ratio of age-specific rates. Nevertheless, use of the pooled population seems to avoid some of the more severe biases that can arise with a completely external standard population. Moreover, the SMR calculated with the pooled groups as standard arises naturally at the first cycle of iteration in one of the numerical procedures for fitting the multiplicative model. These features are illustrated in a cohort analysis of Icelandic breast cancer incidence rates.

(*d*) *Effects of birth cohort on breast cancer incidence in Iceland*

Table 4.1 shows the numbers of female breast cancer cases diagnosed in Iceland during 1910–1971 according to five-year interval and decade of birth (Bjarnason *et al.*, 1974). These data can be considered as arising from a large-scale retrospective cohort study that was made possible by the existence of good records and the fact that all diagnoses in a nearly closed population were made by a small number of pathologists. Also shown are the person-years denominators as estimated from census data and the expected number of cases after fitting of the multiplicative model (4.2). Note that the cells in the lower left- and upper right-hand corners of the table are empty, a consequence of the limited period of case ascertainment. This means that the age distributions of the different birth cohorts are extremely different, and, since the cohort effects are strong also, the age-specific rates for the pooled population will not be proportional to the rates for any particular cohort. Thus, we should not expect that SMRs computed using the pooled population as standard will provide very accurate estimates of the relative risk parameters.

Table 4.1 Observed (O) and expected (E) numbers of female breast cancer cases in Iceland during 1910–1971 by age and year of birth, with approximate person–years (P–Y) at risk[a]

Age group (years)		Year of birth										
		1840–1849	1850–1859	1860–1869	1870–1879	1880–1889	1890–1899	1900–1909	1910–1919	1920–1929	1930–1939	1940–1949
20–24	O						2	—	1	1	1	2
	E						0.42	0.52	0.74	0.85	1.30	3.16
	P–Y						41 380	43 650	49 810	58 105	57 105	76 380
25–29	O						—	2	1	1	5	5
	E						1.10	1.37	1.96	2.27	3.47	3.83
	P–Y						39 615	42 204	48 315	57 685	55 965	33 955
30–34	O					1	1	3	7	12	10	
	E					2.38	3.37	4.22	6.12	7.06	10.84	
	P–Y					29 150	38 430	40 810	47 490	55 720	55 145	
35–39	O					6	11	9	14	20	14	
	E					6.01	8.61	10.84	15.88	18.30	14.36	
	P–Y					27 950	37 375	39 935	46 895	54 980	27 810	
40–44	O				7	14	22	25	29	37		
	E				10.13	12.28	17.72	22.56	33.11	38.21		
	P–Y				25 055	27 040	36 400	39 355	46 280	54 350		
45–49	O				21	11	29	33	57	24		
	E				15.21	18.68	27.03	34.75	51.05	28.29		
	P–Y				24 040	26 290	35 480	38 725	45 595	25 710		
50–54	O			15	8	22	27	38	52			
	E			9.71	15.88	19.25	27.96	36.09	53.41			
	P–Y			22 890	23 095	25 410	34 420	37 725	44 740			
55–59	O			10	15	2	26	47	31			
	E			10.61	17.22	21.43	31.45	40.58	29.70			
	P–Y			21 415	21 870	24 240	33 175	36 345	21 320			
60–64	O		8	11	17	23	31	38				
	E		5.68	10.44	17.01	21.47	32.06	41.34				
	P–Y		17 450	19 765	20 255	22 760	31 965	34 705				
65–69	O		8	10	24	30	53	26				
	E		7.71	14.44	23.67	30.32	46.16	28.71				
	P–Y		15 350	17 720	18 280	20 850	29 600	15 635				
70–74	O	5	3	10	18	22	30					
	E	2.92	5.14	9.74	16.21	21.23	32.77					
	P–Y	9 965	12 850	15 015	15 725	18 345	26 400					
75–79	O	1	7	11	26	32	17					
	E	3.62	6.64	12.80	21.83	28.75	20.37					
	P–Y	8 175	11 020	13 095	14 050	16 480	10 885					
80–84	O	5	8	17	32	31						
	E	4.46	8.85	16.28	31.19	32.23						
	P–Y	7 425	10 810	12 260	14 780	13 600						

[a] From Breslow and Day (1975)

Methods of fitting the multiplicative model by maximum likelihood using the computer program GLIM (Baker & Nelder, 1978) are described below in a more general context. This program uses a modification of the Newton–Raphson algorithm to solve the nonlinear likelihood equations; standard errors of the parameter estimates arise as a by-product of these calculations. For the particular model (4.2), however, there is an alternative fitting algorithm, use of which provides greater insight into the relationship between model fitting and the technique of indirect standardization (Breslow & Day, 1975). The equations that determine the maximum likelihood solution may be written

$$\theta_j = \frac{D_j}{\sum_{k=1}^{K} \psi_k n_{jk}} \quad (j = 1, \ldots, J)$$

and (4.4)

$$\psi_k = \frac{O_k}{\sum_{j=1}^{J} \theta_j n_{jk}} \quad (k = 1, \ldots, K),$$

where $D_j = \sum_k d_{jk}$ are the total deaths at age j and $O_k = \sum_j d_{jk}$ the total deaths at exposure level k (Table 3.4). Inserting initial values $\psi_k^{(0)} = 1$ in the first equation leads to $\theta_j^{(1)} = D_j/N_j$, the marginal death rate in the jth age group, as the initial estimate of θ_j. Here, $N_j = \sum_k n_{jk}$ denotes the total person-years in the jth group. Substituting $\theta_j^{(1)}$ in the second equation gives an initial estimate for ψ_k of $\psi_k^{(1)} = O_k/\sum_j (n_{jk}D_j/N_j)$. Thus, the first-cycle estimate of ψ_k is simply the SMR for the kth exposure group, computed using the age-specific rates for the pooled exposure groups as the standard. Refinements to the initial estimate are obtained by sutstituting $\psi_k^{(1)}$ in the first equation to obtain $\psi_k^{(2)}$, and continuing until convergence when both sets of equations are satisfied simultaneously. If $\hat{\psi}_k$ and $\hat{\theta}_j$ denote the maximum likelihood estimates found at convergence, $\hat{\psi}_k$ may be interpreted as an SMR using the estimated rates $\hat{\theta}_j$ as standard.

Model (4.2) is over-parametrized in the sense that if a particular set of $J + K$ numbers θ_j and ψ_k satisfy the model equation, then so do the sets $\alpha\theta_j$ and $(1/\alpha)\psi_k$ for any positive α. Statisticians refer to such a situation, in which there are more free parameters than can be estimated from the data, as the problem of nonidentifiability. The usual means of solving the problem is to impose constraints on the parameters that are consistent wih a desired interpretation. For the usual choice $\psi_1 = 1$, the remaining ψ_k may be interpreted as relative risks using the first exposure category ($k = 1$) as baseline. The θ_j then correspond to age-specific rates in that baseline category. Of course, the $\hat{\theta}_j$ are actually determined using the data for all the exposure groups, a fact that is especially apparent in this example since for the baseline 1840–1849 cohort data are available for only three age groups.

Another possible resolution of the nonidentifiability problem (Mantel & Stark, 1968) is to choose the normalizing constant α in such a way that when the $\hat{\theta}_j$, interpreted as adjusted age-specific rates, are applied to the pooled population at risk in each age interval, the expected number of deaths is equal to the observed number. Thus,

$$\sum_{j=1}^{J} \hat{\theta}_j N_j = D_+, \tag{4.5}$$

Table 4.2 Results of fitting the multiplicative model to the data in Table 4.1 (10 iterations)[a]

(*a*) *Adjusted SMR by cohort*
Year of birth

1840–	1850–	1860–	1870–	1880–	1890–	1900–	1910–	1920–	1930–	1940–
0.252	0.345	0.558	0.886	0.995	1.067	1.257	1.568	1.541	2.392	4.350

(*b*) *Adjusted age-specific incidence rates per 100 000 person–years*
Age (years)

20–	25–	30–	35–	40–	45–	50–	55–	60–	65–	70–	75–	80–
1.0	2.6	8.2	21.6	45.7	71.4	76.1	88.8	94.8	146.1	116.3	175.3	238.1

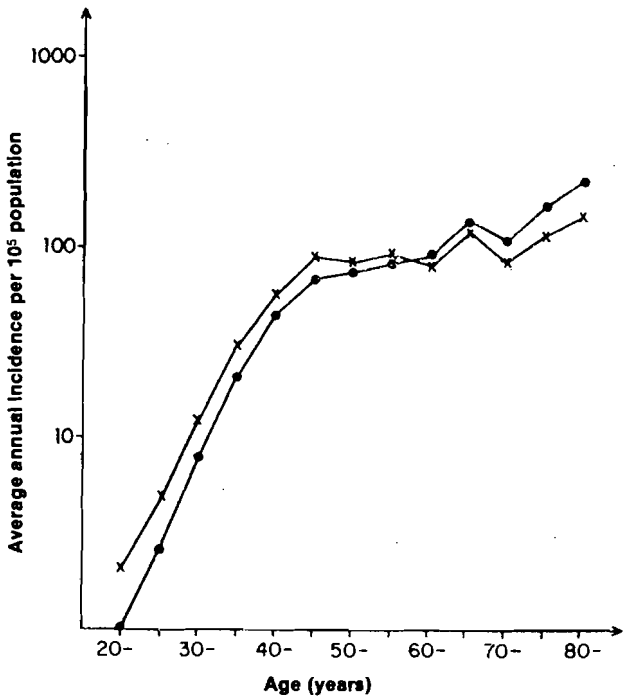
[a] From Breslow and Day (1975)

where $D_+ = \sum D_j$ denotes total deaths. This ensures that the $\hat{\theta}_j$ will be roughly comparable in magnitude to the pooled rates $\hat{\lambda}_j = D_j/N_j$ determined from the marginal totals.

Table 4.2 presents the parameter estimates $\hat{\theta}_j$ and $\hat{\psi}_k$ that arise from fitting model (4.2) under the constraint (4.5). Goodness-of-fit is evaluated by comparing the observed d_{jk} and fitted $\hat{d}_{jk} = \hat{\theta}_j\hat{\psi}_k n_{jk}$ numbers of cases in each cell, both of which are shown in Table 4.1. A summary of the goodness-of-fit is provided by the chi-square statistic

$$\chi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} (d_{jk} - \hat{d}_{jk})^2/\hat{d}_{jk}, \tag{4.6}$$

Fig. 4.2 Crude (×) and fitted (●) age-specific incidence rates for female breast cancer in Iceland, 1911–1972. From Breslow and Day (1975)

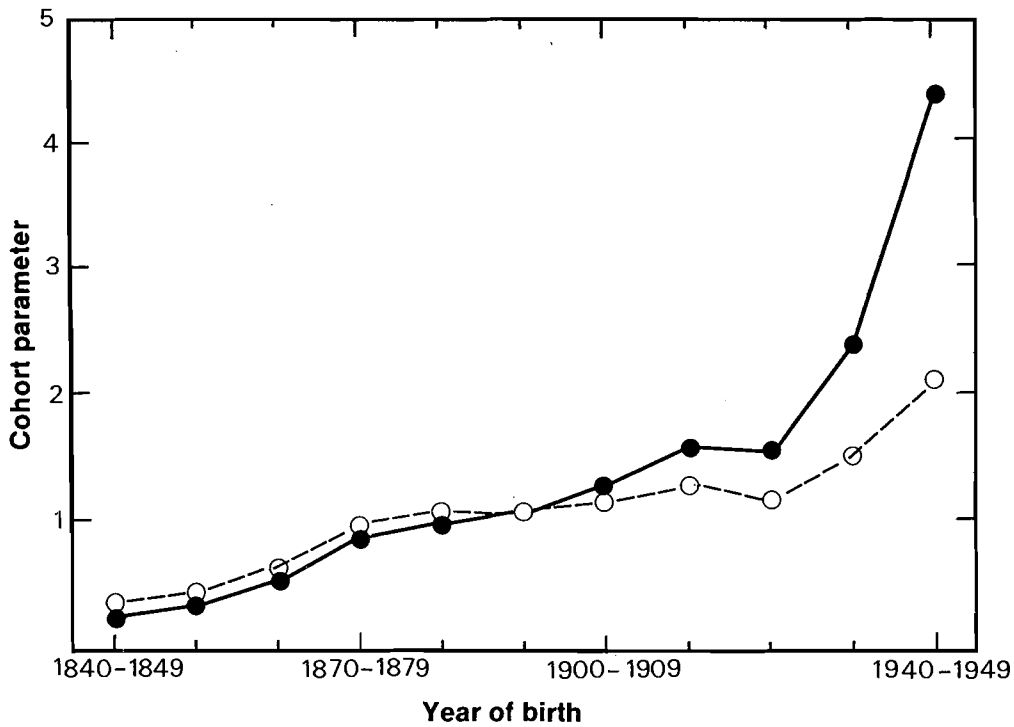


in which the degrees of freedom equal the number of cells with non-zero denominators ($n_{jk} > 0$), minus the number of independently estimated parameters. For our example, (4.6) yields $\chi^2 = 49.0$ with $77 - 23 = 54$ degrees of freedom ($p = 0.67$). It is important that the contributions to chi-square $(d_{jk} - \hat{d}_{jk})^2/\hat{d}_{jk}$ exceed the 95% critical value of 3.84 for a squared normal deviate for only one cell: in the youngest age group in the 1890–1899 cohort there were two cases observed *versus* only 0.42 expected. Thus, the fit appears remarkably good.

The estimates $\hat{\theta}_j$ are plotted on a semilogarithmic scale in Figure 4.2 together with the marginal rates $\hat{\lambda}_j = D_j/N_j$. It is clear that pooling several heterogeneous birth cohorts has overemphasized the change in slope of the age-incidence curve that occurs around the time of the menopause. This is because the marginal rates at older ages are based on earlier birth cohorts which had lower incidence, whereas the marginal rates at younger ages are based on recent cohorts with high incidence. The fitted values $\hat{\theta}_j$ give an impression of the shape of the age relationship for breast cancer that is more comparable to those seen in other populations (Moolgavkar *et al.*, 1980).

A similar disparity between the SMR_ks determined using the marginal rates as standard and the fitted parameters $\hat{\psi}_k$ representing birth cohorts effects is shown in Figure 4.3 (Hoem, 1987). Here, the expected numbers of cases for recent birth cohorts are too high since only marginal rates for young women are used in their calculation, whereas the expected numbers for the earliest cohorts use only the rates at the oldest

Fig. 4.3 Comparison of indirect standardization and multiplicative model fitting in cohort analysis of female breast cancer in Iceland; ○, standardized mortality ratio; ●, multiplicative parameter. From Hoem (1987)



ages. The estimated effect for the 1940–1949 cohort should probably be ignored as it is based on only seven cases occurring at young ages.

4.2 The Poisson assumption[1]

The Poisson model is used throughout this monograph for purposes of making statistical inferences about rates. Specifically, the number of deaths d occurring in a particular age-time-exposure cell is assumed to take on the values $x = 0, 1, 2, \ldots$ with probabilities

$$\text{pr}\,(d = x) = \exp\,(-\lambda n)(\lambda n)^x/x!, \tag{4.7}$$

where λ denotes the unknown rate and n is the person-years denominator. Furthermore, the numbers of deaths occurring in different cells are regarded as statistically independent, even if the same individuals contribute person-years observation time to more than one of them. In this section, we explore the assumptions required for (4.7) to provide a reasonably accurate description of the statistical fluctuations in a collection of rates. Pocock *et al.* (1981) and Breslow (1984a) have developed some alternative models and techniques that may be used in cases in which the observed variation in rates is greater than that predicted by Poisson theory.

(*a*) *Exponential survival times*

Suppose, for simplicity, that there is a single study interval or cell to which I individuals ($i = 1, 2, \ldots, I$) contribute person-years observation times t_i. Set $\delta_i = 1$ if the ith person dies from (or is diagnosed with) the disease of interest in that cell after observation for t_i years; otherwise, $\delta_i = 0$. We further suppose (although this may be unrealistic in certain applications) that there is a fixed maximum time T_i for which the ith individual will be observed if death does not occur. Most frequently, T_i represents the limitation on the period of observation imposed by the person's entry in the middle of the study or his withdrawal from observation at its end (see Fig. 2.1). Thus, $t_i = T_i$ if $\delta_i = 0$, in which case we say that the observation t_i is censored on the right by T_i.

Inferences are to be made about the death rate λ, defined as the instantaneous probability $\lambda\,dt$ that someone dies in the infinitesimal interval $(t, t + dt)$ of time, given that he was alive and under observation at its start. We assume that the rate λ remains constant for the entire period that each individual is under observation. While obviously only an approximation to the true situation, in practice this means that the cell should be constructed to represent a reasonably short interval of age and/or calendar time and that the corresponding exposure category should be fairly homogeneous. Thus, for example, thinking of duration of employment as a measure of exposure, a particular cell might refer to deaths and person-years that occurred between the ages of 55 and 59 during the years 1960–1964 for persons who had been employed for at least 25 and no more than 30 years. We also make the entirely

[1] This section treats a specialized and rather technical topic. Since it presumes greater familiarity with probability theory and statistical inference than the other sections, it may be omitted at first reading.

plausible assumption that the death of one individual has no effect on the outcome for another, or in other words that the two outcomes are statistically independent. Under these conditions the exact distribution of the data (t_i, δ_i) for $i = 1, \ldots, I$ is that of a series of censored exponential survival times.

The exponential distribution has a long history of use in the fields of biometrics, reliability and industrial life testing (Little, 1952; Epstein, 1954; Zelen & Dannemiller, 1961). The ith individual contributes a factor $\lambda e^{-\lambda t_i}$ to the likelihood if he is observed to die during the study interval ($t_i < T_i$, $\delta_i = 1$) and a factor $e^{-\lambda t_i}$(or $e^{-\lambda T_i}$) if he survives until withdrawal ($t_i = T_i$, $\delta_i = 0$). Thus, the log-likelihood function is written

$$L(\lambda) = \sum_{i=1}^{I} (\delta_i \log \lambda - t_i \lambda) = d \log (\lambda) - n\lambda, \tag{4.8}$$

where $d = \sum_i \delta_i$ denotes the total number of events observed and $n = \sum_i t_i$ the total person-years observation time in the specified cell. The elementary estimate $\hat{\lambda} = d/n$ introduced in Chapter 2 is thus seen to be maximum likelihood; it satisfies the likelihood equation $\partial L/\partial \lambda = d/\lambda - n = 0$.

The exact probability distribution of $\hat{\lambda}$ is extremely complicated due to the presence of the censoring times T_i (Kalbfleisch & Prentice, 1980). Mendelhall and Lehman (1960) and Bartholomew (1963) have investigated the first few moments of the distribution, or rather that of the estimated mean survival time $1/\hat{\lambda}$, under the restriction that the censoring times are constant ($T_i = T$ for all i). Approximations to the first two moments are available when the T_i vary. However, these results are all sufficiently complex as to discourage their application to routine problems. One tends to rely instead on large sample normal approximations to the distribution that are based on the log-likelihood (4.8).

(*b*) *The Poisson model*

One reason for the complexity of the exact distribution of d/n is the fact that the observation time is terminated at $t_i < T_i$ for individuals who die. Much simpler distributional properties would obtain if each such subject were immediately replaced by an 'identical' one at the time of death, a type of experimental design that is possible in industrial life testing. For then, considering the ith individual and all subsequent replacements as a single experimental unit, the times of death or failure for that unit constitute observations on a single Poisson process on the interval $0 \leqslant t \leqslant T_i$. The δ_i, which could then take on the values 0, 1, 2, ... rather than just 0 or 1, would have exact Poisson distributions with means λT_i. Since the sum of independent Poisson variables is also Poisson, it follows that the sampling distribution of $d = \sum_i \delta_i$ would be given precisely by (4.7) with $n = \sum_i T_i$, a fixed quantity.

Noting the problems caused by the random observation times t_i, Bartholomew (1963) proposed simply to ignore them in order to obtain an alternative estimate of λ with a more tractable sampling distribution. The only random variables are then the δ_i, which have independent Bernoulli (0/1) distributions with probabilities $p_i = \text{pr}\,(\delta_i = 1) = 1 - \exp(-\lambda T_i)$. If all $T_i = T$, $d = \sum_i \delta_i$ follows the binomial law exactly. If the T_i vary, but either they or λ are sufficiently small that $\text{pr}\,(d = 1)$ is moderate, then an

extension of the usual Poisson approximation to the binomial distribution (Armitage, 1971) shows that d is approximately Poisson with mean $\lambda \sum_i T_i$.

Both lines of reasoning suggest that the Poisson approximation to the exact sampling distribution of d/n, i.e., treating d as Poisson with fixed mean λn as in (4.7), will be adequate, provided that λ is sufficiently small and that only a fraction of the cohort members are expected to become incident cases or deaths during the period in question. Then, the withdrawal of such cases from observation will have a negligible effect on the total observation time and $\sum_i t_i$ will approximate $\sum_i T_i$. From another point of view, the number of 'units' that experience more than one event in the fictitious experiment described above will be negligible. Thus, when the number of deaths or cases d is small in comparison with the total cohort size, a condition which holds for many of the cohort studies of particular cancers that we have in mind, the Poisson model should provide a reasonable approximation to the exact distribution of the rate. Under these same conditions, moreover, the numbers of deaths occurring in different cells may be regarded as statistically independent. Due to their rarity, deaths occurring in one interval will have negligible effects on the probability that a specified number of deaths occurs in the next interval, even though they remove the individuals in question from risk. Hoem (1987) provides a formal statement and proof of this property that is based on unpublished work of Assmussen.

(c) *Asymptotic normality*

If the cases are numerous enough to make up a considerable fraction of the total cohort, the arguments just used to justify the Poisson approximation do not apply. One would probably tend not to use exact Poisson probabilities when the events are numerous anyway, but would instead rely on the approach to normality of estimators and tests based on the Poisson model. This point of view provides some reassurance regarding our reliance on approximate methods of inference based on the likelihood function. Since the log-likelihoods for the Poisson and exponential distributions, both being given by equation (4.8), are identical, it makes no difference which sampling framework we adopt for purposes of making likelihood inferences. The usual large-sample distributions of the maximum likelihood estimates and associated statistics are the same, whether we regard the number of deaths as random and the observation times as fixed, the times as random and the number of deaths as fixed, or both times and number of deaths as random.

This conclusion holds also for problems with multiple cells and rates. Suppose there are J cells with associated death rates λ_j, and let δ_{ij} denote whether ($\delta_{ij} = 1$) or not ($\delta_{ij} = 0$) the ith individual dies in the jth cell, while t_{ij} denotes his contribution to the observation time n_j in that cell. According to the general theory of survival distributions (Kalbfleisch & Prentice, 1980; see also §5.2), the log-likelihood of the data may be written

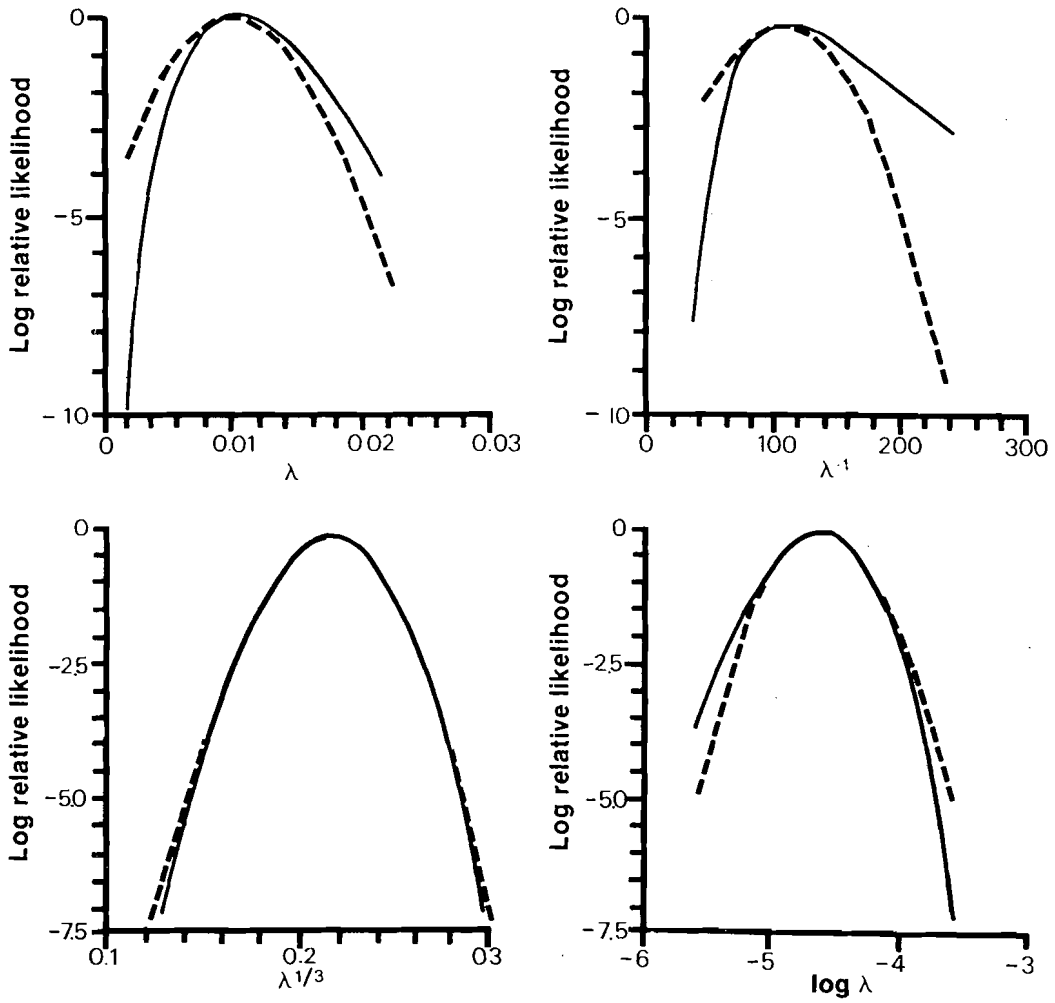
$$L(\boldsymbol{\lambda}) = L(\lambda_1, \ldots, \lambda_J) = \sum_{i=1}^{I} \sum_{j=1}^{J} \delta_{ij} \log \lambda_j - t_{ij}\lambda_j = \sum_{j=1}^{J} d_j \log \lambda_j - n_j\lambda_j. \qquad (4.9)$$

This likelihood also arises when the d_j are independent Poisson variables with means

$n_j\lambda_j$ or when the t_{ij} form a censored sample of independent exponential survival times with death rate parameters λ_j (Holford, 1980). However, because of the dependencies between deaths that occur in different intervals and the fact that the n_j are random variables, neither of these exact sampling models is strictly correct. While they are adequate for large-sample likelihood inferences, such as made in this section, other properties based on the Poisson model (such as the standard errors given by equations (2.6) and (2.7)) may require also that the deaths be only a small fraction of the total persons in each cell in order that these be reasonably accurate (Hoem, 1987).

In the sequel, log-likelihood functions similar to (4.9) will be considered as functions of a relatively small number of unknown parameters that describe the structure in the rates. The shape of the log-likelihood function can change drastically depending upon the model selected or even upon the choice of parameters used to describe a given model. As an example, suppose that ten deaths are observed in a single cell with 1000 person-years of observation. Figure 4.4 contrasts the shape of the log-likelihood (4.9) of these data considered as a function of: (i) the death rate λ itself; (ii) the expected

Fig. 4.4 Exact (—) and approximate (----) log-likelihoods for various parametrizations of the death rate, λ, when $d = 10$ and $n = 1000$



lifetime $1/\lambda$; (iii) the cube-root transform $\lambda^{1/3}$, and (iv) the log death rate $\log \lambda$. Also shown are quadratic approximations to each likelihood that would apply if the data were normally distributed with a mean value equal to the unknown parameter and a fixed variance given by the observed information[1] function evaluated at the maximum likelihood estimate. Comparison of the four figures shows that the cube-root and log parametrizations yield the most 'normal' looking likelihoods, whereas those for λ and especially $1/\lambda$ are rather skewed. The cube-root transform, which also occurred in Byar's approximation to Poisson error probabilities (equations (2.11) and (2.13)), has the property that it exactly eliminates the cubic term in a series expansion of L about the maximum likelihood estimate (Sprott, 1973). Empirical work by Schou and Vaeth (1980) has confirmed that the sampling distributions of $\log \hat{\lambda}$ and $\hat{\lambda}^{1/3}$ are more nearly normal in finite samples than those of $\hat{\lambda}$ or its inverse.

The implication of these results for the statistician is that statistical inferences that rely on asymptotic normal theory are better carried out using procedures that are invariant under transformations of the basic parameters. The maximum likelihood estimate itself satisfies this requirement, as do likelihood ratio tests, score tests computed with expected information, and confidence intervals obtained by inverting such invariant tests. However, procedures based on a comparison of the point estimate with its standard error as obtained from the normal (quadratic) approximation to the log-likelihood are not generally reliable and should be used only if the normal approximation is known to be good (Vaeth, 1985). This condition is met for parameters in the standard multiplicative models considered below, as it was for the logistic models discussed in Volume 1. It is not met for other models, as we shall see.

4.3 Fitting the multiplicative model

Most of the features of the multiplicative model for rates are already present in the two-dimensional table considered in §4.1. We continue to think of the basic data as being stratified in two dimensions. The first dimension corresponds to nuisance factors such as age and calendar time, the effects of which on the baseline rates are conceded in advance and are generally of secondary interest in the study at hand. The second dimension corresponds to the exposure variables, the effects of which we wish to model explicitly. The total number of cells into which the data are grouped is thus the product of J strata and K exposure categories. The basic data consist of the counts of deaths d_{jk} and the person-years denominators n_{jk} in each cell, together with p-dimensional row vectors $\mathbf{x}_{jk} = (x_{jk}^{(1)}, \ldots, x_{jk}^{(p)})$ of regression variables. These latter may represent either qualitative or quantitative effects of the exposures on the stratum-specific rates, interactions among the exposures and interactions between exposure variables and stratification (nuisance) variables.

[1] Recall from §6.4 of Volume 1 or elsewhere that the information is defined as minus the second derivative of L. Since we consider some models in this volume for which the information depends on the data, a distinction is made between the observed information and its expectation. The latter is also known as Fisher information.

(a) The model equation

A general form of the multiplicative model is

$$\log \lambda_{jk} = \alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta}, \tag{4.10}$$

where the λ_{jk} are the unknown true disease rates, the α_j are nuisance parameters specifying the effects of age and other stratification variables, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional column vector of regression coefficients that describe the effects of primary interest. An important feature of this and other models introduced below is that the disease rates depend on the exposures only through the quantity $\alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta}$, which is known as the linear predictor. If the regression variables $\mathbf{x}_{jk}$ depend only on the exposure category k and not on j, (4.10) specifies a purely multiplicative relationship such that the ratio of disease rates $\lambda_{jk}/\lambda_{jk'}$ for two exposure levels k and k', namely $\exp\{(\mathbf{x}_k - \mathbf{x}_{k'})\boldsymbol{\beta}\}$, is constant over the strata. Evaluation of the goodness-of-fit of such models informs us as to whether a summary of the data in terms of relative risk is reasonably plausible. If the ratios $\lambda_{jk}/\lambda_{jk'}$ change with j, additional variables x_{jk} which depend on both j and k and describe interactions between stratum and exposure effects may be needed to provide a comprehensive summary of the data.

The simple multiplicative model (4.2) for the two-dimensional table of rates is expressed by taking the x variables to be dummy or indicator variables with a value of 1 for a particular exposure category and 0 elsewhere. A total of $K-1$ such indicator variables is needed to express the relative risks associated with the different exposure categories, the first level ($k=1$) typically being used as a reference or baseline category. The advantage of the more general model (4.10) is that it allows us to quantify the relative risks according to measured dose levels, impose some structure on the joint effects of two or more exposures, and relax the strict multiplicative hypothesis through the introduction of interaction terms. These features are developed below in a series of examples. However, we first discuss implementation of the methodology using the Royal Statistical Society's GLIM program for fitting generalized linear models (Baker & Nelder, 1978).

(b) Fitting the model with GLIM

Input to GLIM or other standard programs will consist of up to JK data records containing the counts d_{jk} of disease cases or deaths, the person-years denominators n_{jk}, the values $x_{jk}^{(1)}, \ldots, x_{jk}^{(p)}$ of the regression variables to be included in the model, and sufficient additional data to identify each stratum (j) and exposure category (k). Records for (j, k) cells with no person-years of observation ($n_{jk}=0$) are usually omitted.

If some or all of the exposures are to be analysed as qualitative or discrete variables, it is not necessary to construct the 0/1 indicators explicitly for each exposure category or stratum, since GLIM makes provision in its FACTOR command for designating certain input variables as qualitative. Their values $(1, 2 \ldots)$ are then presumed to designate the factor level. By default, the first level is taken as baseline, binary indicator variables being constructed by the program for each higher level.

According to §4.2 the numbers of deaths d_{jk} from a specific cause may be regarded

as independent Poisson variables with mean values $E(d_{jk}) = n_{jk}\lambda_{jk}$. In view of (4.10) we have

$$\log E(d_{jk}) = \log (n_{jk}) + \alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta}. \tag{4.11}$$

Since the log transform of the mean is a linear function of the unknown parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the model conforms to the usual log-linear model for Poisson variables and as such is easily fitted using standard features of GLIM. Note that the constants $\log (n_{jk})$ offset the model equation (4.11) from the origin in the sense that the log mean equals $\log (n_{jk})$ when the α and β parameters are zero. This means that a variable containing the log person-years denominators is declared an OFFSET when invoking the program. In order to fit a separate α_j for each stratum level, it is easiest to create a stratum variable taking values $j = 1, \ldots, J$ and declare it as a FACTOR. When the strata are formed by combinations of two or more variables, these may each be declared FACTORs and included in the model with all their interactions. The exposures are treated as either variables or factors, depending upon whether quantitative or qualitative (categorical) effects are to be specified.

An alternative GLIM approach is to define the dependent or y variable as the observed rate $\hat{\lambda}_{jk} = d_{jk}/n_{jk}$ and to declare the person-years denominator n_{jk} as a prior WEIGHT. Then, no OFFSET is needed. This approach also applies with the additive (identity) and power 'link' functions considered in §4.4, whereas the approach that declares $\log (n_{jk})$ to be an OFFSET does not. See Frome and Checkoway (1985).

(c) *Summary measures of fit*

Summary measures of fit give an overall evaluation of the agreement between observed and fitted values. Two are in common use. One is the χ^2 statistic already defined in (4.6) as the sum of the squared residuals. The other is the log-likelihood ratio statistic that compares the observed and fitted values *via*

$$G^2 = 2\left\{\sum_{j=1}^{J}\sum_{k=1}^{K} d_{jk} \log (d_{jk}/\hat{d}_{jk}) + (\hat{d}_{jk} - d_{jk})\right\}. \tag{4.12}$$

G^2 is known as the deviance in GLIM. G^2 and χ^2 are referred to tables of the chi-square distribution in order to ascertain the overall goodness-of-fit. They tend to give similar values in most applications. Both may overstate the degree of departure from the fitted model when many cells contain small counts (Fienberg, 1980) and when they are interpreted as chi-square statistics; a correction factor for G^2 is available (Williams, 1976).

The degrees of freedom associated with these statistics equal the number of cells with nonzero person-years of observation minus the number of linearly independent parameters in the model, namely $J + p$ in the above formulation. When the value of G^2 or χ^2 exceeds its degrees of freedom by an amount significantly greater than expected under chi-square sampling, we conclude that the fit is inadequate. Either there are systematic effects that have not been accounted for by the model, or else the random variation in disease rates among neighbouring cells is greater than that specified by the Poisson assumption. Agreement between the deviance and its degrees of freedom does

not guarantee that the fit is good, however, particularly when the degrees of freedom are large. Systematic patterns or trends in the residuals that may be indicative of departures from model assumptions, and large residual values for individual cells, often are not reflected adequately in the summary measure. Also, a good fit for a model based on a cross-classification that ignores relevant covariables does not imply that such variables are unimportant or should not be considered.

(*d*) *Adding variables to the model equation*

The most common remedy for the lack of fit of a given model equation, or for examining whether systematic departures from model assumptions are being obscured by the global goodness-of-fit statistic, is to add regression variables. Indeed, the process of model building generally involves fitting a hierarchy of model equations that represent increasing degrees of complexity in the relationship between the relative risk and the exposure variables, or increasingly complex interaction (modifying) effects of the stratification variables with the exposure variables. Comparison of the goodness-of-fit measures for two different models, one of which is contained within the other, provides a formal test of the statistical significance of the additional variables. Thus, if G_1^2 and G_2^2 are the deviances for models 1 and 2, where model 2 contains q more independent parameters than model 1, the difference $G_1^2 - G_2^2$ is treated as a chi-square statistic with q degrees of freedom for testing the significance of the additional variables. Two other commonly used tests, one based on the estimated regression coefficients and the other on the efficient score (first derivative of the log-likelihood), are briefly described in §6.4 of Volume 1.

(*e*) *Further evaluation of goodness-of-fit: analysis of residuals*

The extent to which the model summarizes the data can be evaluated globally by an overall goodness-of-fit test, but often a more informative approach is to examine how well the number of deaths in each cell is predicted. This is accomplished by comparing the observed numbers of deaths d_{jk} in each cell with the fitted number $\hat{d}_{jk} = n_{jk} \exp(\hat{\alpha}_j + \mathbf{x}_{jk}\hat{\beta})$, where $\hat{\alpha}_j$ and $\hat{\beta}$ denote the maximum likelihood estimates. In order to get some idea of whether the deviations between observed and fitted values are greater than would be expected from sampling (Poisson) variability, we calculate the standardized residuals $r_{jk} = (d_{jk} - \hat{d}_{jk})/\sqrt{\hat{d}_{jk}}$. Since they have the form of the difference between an observation and its estimated mean, divided by the estimated standard deviation under the Poisson model, the r_{jk} may be regarded roughly as equivalent normal deviates when assessing the fit for any particular cell. A refinement, taking account of the number of fitted parameters, is to consider as equivalent normal deviates the adjusted residuals

$$\tilde{r}_{jk} = \frac{r_{jk}}{(1-h_{jk})^{1/2}} = \frac{d_{jk} - \hat{d}_{jk}}{\{\hat{d}_{jk}(1-h_{jk})\}^{1/2}}, \tag{4.13}$$

where the h_{jk} denote the diagonal element of the 'hat' or projection matrix that arises in the theory of linear regression (Hoaglin & Welsh, 1978). These are available in

GLIM as the product of the 'iterative weights', which equal $\hat{d}_{jk}$ for the multiplicative Poisson model, times the variances of the linear predictors. In other words, for the multiplicative model,

$$h_{jk} = \hat{d}_{jk} \operatorname{Var} (\hat{\alpha}_j + \mathbf{x}_{jk}\hat{\boldsymbol{\beta}}). \tag{4.14}$$

The sum of the h_{jk} equals the number of parameters estimated, namely $J + p$.

Modern texts on regression anslysis (e.g., Cook & Weisberg, 1982) devote considerable attention to graphical methods of residual analysis. Certain patterns in the residuals are indicative of specific types of departures from model assumptions. For example, a tendency for the absolute values $|r_{jk}|$ to increase with $\hat{d}_{jk}$ would indicate that the equality of mean and variance specified by the Poisson model was inadequate and that the variability increased faster than as a linear function of the mean. Correlations between the residuals and regression variables not yet included in the model equation would indicate that the model was incomplete, whereas correlations with certain functions of the fitted values may indicate that the log-linear specification (4.11) is inadequate and that the death rates λ_{jk} are better modelled by some other function of the linear predictors (Pregibon, 1980). We present some examples of graphical residual analyses in the sequel, but systematic discussion of their rationale and use is beyond the scope of this monograph.

(*f*) *Gauging the influence of individual data points*

Another aspect of model checking, apart from examination of residuals, is to determine the influence that individual data points have on the estimated regression coefficients $\hat{\alpha}_j$ and $\hat{\boldsymbol{\beta}}$. The investigator needs to be aware whenever elimination of one of the (j, k) cells from the analysis would lead to a particularly marked change in the fitted model. Sometimes, such influential cells are also 'outliers', in the sense that the multivariable observation $(d_{jk}, n_{jk}, \mathbf{x}_{jk})$ is far removed from the rest of the data. It is important to check that such data have been correctly recorded and are not in error. The same is true for data points that give rise to large residuals. 'Robust' regression methods have been developed specifically to reduce the influence of such outlying observations (Huber, 1983); however, the rationale for their use is not entirely clear when the data in question are known to be valid. A concerted effort to understand why the particular observation does not conform to the rest of the data may be more important than finding the model that best fits when that point is removed.

Influential data points are often reasonably well fitted by the model and not amenable to detection by an examination of their residuals. More sensitive measures of influence can be developed using a combination of the residuals and the diagonal elements h_{jk} of the 'hat' matrix (equation 4.14). A rough rule of thumb for general applications is to regard an individual observation as having a particularly heavy influence on the overall fit if the corresponding h_{jk} exceeds twice the average value (Hoaglin & Welsh, 1978). This rule is not applicable in the present context, however, since cells with large person-years and expected numbers of cases will necessarily have a large impact on the fit. Rather, we use the h_{jk} diagnostics in a descriptive and comparative manner to identify those cells that have the greatest overall influence on

the fit and to demonstrate that the relative influence of different cells on the regression coefficients can depend on the transformation linking the rates λ_{jk} to the linear predictor.

Measures of the influence of individual cells on particular regression coefficients involve these same basic quantities (Pregibon, 1979, 1981). In particular, an approximation to the change in the estimated regression coefficients $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ that is occasioned by deletion of the (j, k) cell from the statistical analysis is given by

$$\Delta(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})_{-jk} \simeq -\hat{\boldsymbol{\Sigma}}\, \mathbf{x}^*_{jk}(d_{jk} - \hat{d}_{jk})/(1 - \hat{h}_{jk}), \tag{4.15}$$

where $\hat{\boldsymbol{\Sigma}}$ denotes the asymptotic covariance matrix of the estimates $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ and $\mathbf{x}^*_{jk} = (0, \ldots 1, \ldots, 0, \mathbf{x}_{jk})$ denotes an augmented vector of regression variables preceded by J stratum indicators of which the jth equals one.

Example 4.1

Appendix VI contains grouped data from a recent update (Peto, J. *et al.*, 1984) of the Welsh nickel refinery workers study that is described in detail in Appendix ID. Previously published data from this study (Doll *et al.*, 1970) were used in §3.5 to illustrate techniques of internal standardization. The latest follow-up through 1981 uncovered 137 lung cancer deaths among men aged 40–85 years and 56 deaths from cancer of the nasal sinus.

Nasal sinus cancer deaths and person-years of observation are classified in Appendix VI by three risk factors: (i) age at first employment (AFE) in four levels (1 = <20; 2 = 20–27.4; 3 = 27.5–34.9; and 4 = 35+ years); (ii) calendar year of first employment (YFE) in four levels (1 = <1910; 2 = 1910–1914; 3 = 1915–1919; and 4 = 1920–1924); and (iii) time since first employment (TFE) in five levels (1 = 0–19; 2 = 20–29; 3 = 30–39; 4 = 40–49; and 5 = 50+ years). Since less than one case of nasal sinus cancer would have been expected from national rates, it was deemed unnecessary to account for the background rates. Instead, the object was to study the evolution of nasal sinus cancer risk as a function of time since first exposure, and to determine whether this was influenced by the age and year in which that exposure began.

Table 4.3 displays the GLIM commands needed to read the 72 data records, fit the log-linear model with main effects for factors AFE, YFE and TFE, and print the results shown in Tables 4.5 and 4.6. Models involving a number of other combinations of these same factors were investigated also. Their deviances, displayed in Table 4.4, demonstrate that the three factors have strong, independent effects on rates of nasal sinus cancer. The log-likelihood ratio statistics of $95.6 - 58.2 = 37.4$ for AFE, $83.5 - 58.2 = 25.3$ for YFE and $70.8 - 58.2 = 12.6$ for TFE, with 3, 3 and 4 degrees of freedom, are all highly significant. The parameter estimates in Table 4.5 indicate that nasal sinus cancer risk increases steadily with both age at and time since first exposure, and that it peaks for men who were first employed in the 1910–1914 period. Since the global tests for two-factor interactions are of at most borderline significance, the largest being 16.4 (9 degrees of freedom, $p = 0.06$) for YFE × TFE, we conclude that the simple multiplicative model provides a reasonable description of the data. Further support for this conclusion is obtained by comparing observed and fitted numbers of cases classified by AFE × TFE collapsing over YFE (Table 4.6), and similarly for the other two-factor combinations. The greatest discrepancy is observed for the YFE × TFE cross-classification (not shown), where four cases are observed in the cell with YFE = <1910 and TFE = 20–29 years, whereas only 1.30 are expected under the model ($\chi^2_1 = 5.6$). We are inclined to interpret this aberrant value as a chance occurrence.

The marginal totals of expected numbers of deaths in Table 4.6 agree exactly with the observed numbers, which confirms this as a defining characteristic of the maximum likelihood fitting of the log-linear model (Fienberg, 1980). Inclusion of the main effects of AFE, YFE and TFE in the model ensures that the fitted values for each of these factors, when summed over the levels of the other two, will agree with the subtotals of observed values. (Inclusion of the AFE × TFE interactions in the model in addition to the main effects would result in subtotals of fitted values for the AFE × TFE two-dimensional marginal table that agree with the corresponding observed subtotals.) Table 4.6 also illustrates a fundamental property of the 'hat' matrix elements, h, namely, that their grand total equals the number of independent parameters in the model. In this example, there is one parameter associated with the constant term or grand mean (see Table 4.5), three

Table 4.3 GLIM commands used to analyse the data in Appendix VI

```
$UNITS 72 ! 72 DATA RECORDS IN FILE IN APPENDIX VI; EQUATE TO FORTRAN UNIT 1
$DATA AFE YFE TFE CASE PY! NAMES OF 5 VARIABLES TO BE READ FROM FILE
$DINPUT 1 80 ! READ DATA FROM FORTRAN UNIT 1
$FACTOR 72 AFE 4 YFE 4 TFE 5 ! DECLARE FACTORS WITH 4 AND 5 LEVELS EACH
$CAL LPY = %LOG(PY) ! CALCULATE LOG PERSON–YEARS
$OFFSET LPY ! DECLARE LOG PERSON–YEARS AS OFFSET TO MODEL EQUATION
$ERR P ! POISSON MODEL WITH DEFAULT (LOG-LINEAR) LINK
$YVAR CASE ! NO. OF NASAL CANCERS (CASE) AS DEPENDENT VARIABLE
$FIT AFE + YFE + TFE ! FIT LOG-LINEAR MODEL WITH MAIN EFFECTS FOR EACH FACTOR
$ACC 5 ! CHANGE NO. OF DECIMALS IN PRINTOUT
$REC 10 $FIT . ! REFIT SAME MODEL FOR GREATER ACCURACY
$DIS M E$ ! DISPLAY MODEL AND PARAMETER ESTIMATES. SEE TABLE 4.5
$EXT %VL %PE ! EXTRACT VARIANCE OF LINEAR PREDICTOR AND PARAMETER ESTIMATES
$VAR 11 PR ! DECLARE REL RISK RR AS VARIABLE OF DIMENSION 11
$CAL RR = %EXP(%PE) $LOOK RR $ ! CALCULATE AND PRINT REL RISKS FOR TABLE 4.5
$CAL H = %WT*%VL ! CALCULATE DIAGONAL ELEMENTS OF 'HAT' MATRIX H
$CAL I = 5*(AFE-1) + TFE ! SET UP INDEX FOR CELLS IN AFE BY TFE MARGINAL TABLE
$VAR 20 CAST EXPT PYT HT ! SET UP VARIABLES OF DIMENSION 20
$CAL CAST = 0 : EXPT = 0 : PYT = 0 : HT = 0 ! INITIALIZE ARRAYS
$CAL CAST(I) = CAST (I) + CASE : EXPT(I) = EXPT(I) + %FV : PYT(I) + PYT(I) + PY$
$CAL HT(I) = HT(I) + H ! CULULATE SUBTOTALS OF CASES, FITTED VALUES ETC. OVER YFE
$LOOK CAST EXPT PTY HT ! PRINTOUT FOR TABLE 4.6
$STOP
```

Table 4.4 Goodness-of-fit statistics (deviances) for a number of multiplicative models fitted to the data on Welsh nickel refinery workers in Appendix VI

Factors in model[a]	Degrees of freedom	Deviance
—	71	135.7
AFE	68	109.1
YFE	68	100.6
TFE	67	120.6
AFE + YFE	65	70.8
AFE + TFE	64	83.5
YFE + TFE	64	95.6
AFE + YFE + TFE	61	58.2
AFE*YFE[b] + TFE	52	49.2
AFE*TFE + YFE	50	48.5
YFE*TFE + AFE	50	41.8

[a] AFE, age at first employment; YFE, year of first employment; TFE, time since first employment

[b] AFE*YFE indicates, in standard GLIM notation, that both main effects and first-order interactions involving the indicated factors are included in the model equation.

Table 4.5 Regression coefficients, standard errors and associated relative risks for the multiplicative model fitted to data on nasal sinus cancers in Welsh nickel refinery workers (Appendix VI)

Factor[a]	Level	Regression coefficient ± standard error	Relative risk[b]
AFE	<20	—	1.0
	20.0–27.4	1.67 ± 0.75	5.3
	27.5–34.9	2.48 ± 0.76	12.0
	35+	3.43 ± 0.78	30.8
YFE	<1910	—	1.0
	1910–14	0.62 ± 0.37	1.9
	1915–19	0.05 ± 0.47	1.1
	1920–24	−1.13 ± 0.45	0.3
TFE	<20	—	1.0
	20–29	1.60 ± 1.05	4.9
	30–39	1.75 ± 1.06	5.8
	40–49	2.35 ± 1.07	10.5
	50+	2.82 ± 1.12	16.7
Constant term		−9.27 ± 1.32	Estimated baseline[c] rate of nasal sinus cancer deaths: 9.42 per 100 000 person–years
Deviance: $G^2 = 58.2$ on 61 degrees of freedom			

[a] AFE, age at first employment; YFE, year of first employment; TFE, time since first employment
[b] Exponentiated regression coefficients
[c] For AFE < 20, YFE < 1910 and TFE < 20

each with AFE and YFE and four with TFE, for a total of 11. Note that the larger values of h_+ are generally associated with the cells with the largest number of observed deaths.

4.4 Choosing between additive and multiplicative models

If a good fit is obtainable with the multiplicative model only by introducing complicated interaction terms involving baseline and exposure factors, re-examination of the basic multiplicative relationship is usually in order. It may be that the effects of exposure are better and more easily expressed on another scale. Formal evaluation of the relative merits of the multiplicative and additive models for any particular set of regression variables is made possible by embedding them in a wider class of models that contain both as special cases. One useful class of models for this purpose is the power family

$$\lambda_{jk}^{\rho} = \alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta} \tag{4.16}$$

that relates the disease rates to the linear predictors $\alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta}$ by means of the power transform with exponent ρ (Aranda–Ordaz, 1983). The additive model corresponds to the case $\rho = 1$, whereas, since $(\lambda^{\rho} - 1)/\rho$ tends to $\log \lambda$ in the limit as ρ tends towards zero, the multiplicative model corresponds to $\rho = 0$.

Power models may be fitted easily using GLIM. The dependent or y observations, assumed to have a Poisson error structure, are the rates $\hat{\lambda}_{jk} = d_{jk}/n_{jk}$ rather than

Table 4.6 Results of fitting the multiplicative model to the data on Welsh nickel refinery workers in Appendix VI: observed (O) and expected (E) numbers of nasal sinus cancer deaths, person–years (P–Y) and summed regression diagnostics h by age at first employment and time since first employment

Age at first employment (years)		Years since first employment					
		0–19	20–29	30–39	40–49	50+	Total
<20	O	0	1	0	1	0	2
	E[a]	0.02	0.34	0.43	0.58	0.63	2.00
	P–Y	353.0	806.6	832.4	652.1	445.0	3089.1
	h_+[b]	0.03	0.22	0.26	0.36	0.41	1.28
20.0–27.4	O	0	3	6	7	4	20
	E	0.21	4.40	5.61	5.97	3.80	20.00
	P–Y	1107.9	2044.2	2094.9	1281.8	536.1	7064.9
	h_+	0.23	0.69	0.85	0.94	0.91	3.62
27.5–34.4	O	0	8	5	5	2	20
	E	0.31	6.30	6.62	5.19	1.57	20.00
	P–Y	732.7	1303.8	1098.6	481.4	95.3	3711.8
	h_+	0.34	0.89	0.94	0.86	0.43	3.46
35+	O	1	7	6	0	—	14
	E	0.46	7.96	4.33	1.25	—	14.00
	P–Y	392.4	622.9	303.4	46.1	—	1364.8
	h_+	0.49	1.10	0.78	0.28	—	2.65
Total	O	1	19	17	13	6	56
	E	1.00	19.00	17.00	13.00	6.00	56.00
	P–Y	2586.0	4777.5	4329.3	2461.4	1076.4	15 230.6
	h_+	1.09	2.90	2.83	2.44	1.75	11.00

[a] Expected values adjusted also for year of first employment

[b] Regression diagnostics h summed over levels of year of first employment. These values should *not* be substituted in the expression for adjusted residuals (equation 4.13).

the numbers of deaths; the person-years denominators are treated as prior weights, using the WEIGHT command. The model (4.16) is available as an alternative GLIM 'link' for Poisson observations.

Diagonal elements of the 'hat' matrix are obtained at convergence as

$$h_{jk} = n_{jk} w_{jk} \operatorname{Var}(\hat{\eta}_{jk}),$$

where w_{jk} is the GLIM iterated weight for the power model and $\hat{\eta}_{jk} = \hat{\alpha}_{jk} + \mathbf{x}_{jk}\hat{\boldsymbol{\beta}}$ is the linear predictor. The approximate change in the regression coefficients upon deletion of the (j, k)th cell of data is given by

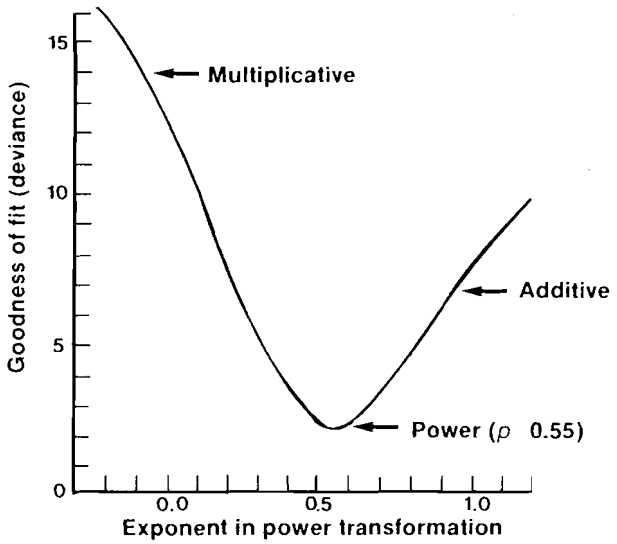
$$\Delta(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})_{-jk} = -\hat{\boldsymbol{\Sigma}}\, \mathbf{x}^*_{jk}\, n_{jk} w_{jk} (y_{jk} - \hat{\eta}_{jk})/(1 - h_{jk}), \tag{4.17}$$

where $\mathbf{x}^*_{jk}$ is the vector of augmented regression variables, y_{jk} denotes the GLIM 'working variable', and $\hat{\boldsymbol{\Sigma}}$ is again the covariance matrix of the estimated parameters.

Example 4.2

The data on coronary deaths among British male doctors shown in Table 3.15 offer a simple example for examining some of these issues regarding goodness-of-fit and model selection. The rate ratios for smokers *versus* nonsmokers decrease with advancing age, while the rate differences generally increase. This suggests

Fig. 4.5 Goodness-of-fit statistics (G^2) for a variety of power models fitted to the data in Table 3.15



that neither the multiplicative nor the additive model is completely appropriate for expressing the effect of smoking in a single number, and that some intermediate power model might work better. Accordingly, several models of the form (4.16) were fitted with a single binary exposure variable coded 0 for nonsmokers and 1 for smokers. Figure 4.5 shows that the minimum value of the deviance G^2, nominally a chi-square-distributed statistic with $10-6=4$ degrees of freedom, occurs in the vicinity of $\rho=0.55$, intermediate between the additive and multiplicative models. Neither of these extremes provides a satisfactory fit, since one finds $G^2=12.1$ for the multiplicative and $G^2=7.4$ for the additive structure, compared with $G^2=2.1$ for the best power transform.

Table 4.7 presents estimates and standard errors for the five α_j parameters and the single smoking coefficient β under each of the three models. When suitably transformed, the α's represent the fitted death rates among nonsmokers per 1000 person-years of observation. Under the additive model, for example, the fitted rate for men aged 55–64 years is 6.2 deaths per 1000 population per year. For the power model the rate per 1000 person-years is $(2.180)^{(1/0.55)}=4.1$, and for the multiplicative model it is $\exp(1.616)=5.0$. Smoking is estimated to increase (add to) the death rate by 0.59 deaths per 1000 person-years under the additive model, while under the multiplicative model smoking multiplies the death rate by $\exp(0.355)=1.43$ at all ages. The smoking effect is not so conveniently expressed on the power scale, but since ρ and β are each about equal to 0.5 it may be roughly described as increasing the square root of the death rate per 1000 person-years by one-half. Note that t statistics of the form $t=\hat{\beta}/SE(\hat{\beta})$ yield roughly comparable values for all three models with these data. The likelihood ratio (deviance) tests for smoking ($\beta=0$) are obtained by subtracting the goodness-of-fit deviances in Table 4.7 from the deviance for the model with age effects only, which equals 23.99 regardless of the value of ρ. Best agreement between the t test and deviance test of smoking effect is found for the multiplicative model.

Table 4.7 also shows the fitted numbers of deaths $\hat{d}_{jk}$ for smokers and nonsmokers under the three models. These were combined with the diagnostic values h_{jk} to calculate adjusted residuals $\tilde{r}_{jk}$ (equation 4.13). The 'hat' matrix elements h_{jk}, approximate changes in the smoking coefficient from equation 4.15 or 4.17 and adjusted residuals $\tilde{r}_{jk}$ are all displayed in Table 4.8.

Examination of the entries in the first two parts of Table 4.8 shows that the data for the youngest age group, which contains a small number of deaths observed in a rather large population, have the greatest influence on the estimated rate difference in the additive model. More of the information about the rate ratio under the multiplicative model comes from older age groups where there are larger numbers of deaths. The power model occupies an intermediate position *vis-à-vis* the diagnostics h_{jk}, and deletion of single cells has little effect on the smoking parameter. The residual patterns are in the anticipated direction, the death rates

Table 4.7 Parameter estimates and fitted values for three statistical models for the data on coronary deaths among British male doctors in Table 3.15

Age range (years)	Parameter	Statistical model[a]		
		Additive ($\rho = 1$)	Power ($\rho = 0.55$)	Multiplicative ($\rho = 0$)
		Parameter estimates ± SE[b]		
35–44	α_1	0.084 ± 0.066	0.276 ± 0.092	−1.012 ± 0.192
45–54	α_2	1.641 ± 0.218	1.115 ± 0.110	0.472 ± 0.130
55–64	α_3	6.304 ± 0.456	2.456 ± 0.132	1.616 ± 0.115
65–74	α_4	13.524 ± 0.964	3.859 ± 0.180	2.338 ± 0.116
75–84	α_5	19.170 ± 1.704	4.763 ± 0.257	2.688 ± 0.125
Smoking	β	0.591 ± 0.125	0.493 ± 0.098	0.355 ± 0.107
$t^2 = (\hat{\beta}/\mathrm{SE}(\hat{\beta}))^2$		19.3	25.3	11.0

	Fitted values[c]					
	Non-smokers	Smokers	Non-smokers	Smokers	Non-smokers	Smokers
35–44	1.59	35.37	1.81	32.53	6.83	27.17
45–54	17.51	96.50	13.00	102.56	17.12	98.88
55–64	35.99	197.25	29.26	204.49	28.74	205.26
65–74	34.96	178.73	30.12	183.61	26.81	187.19
75–84	28.03	105.07	24.97	108.65	21.51	111.49
Goodness-of-fit (deviance) χ^2_4	7.43		2.14		12.13	
Deviance test for smoking effect χ^2_1	16.56		21.85		11.86	

[a] Exponent (ρ) of power function relating death rates and linear predictor
[b] Person–years denominators expressed in units of 1000
[c] See Table 3.15 for observed values

for smokers being seriously underestimated by the multiplicative model in the youngest age group and seriously overestimated in the oldest group. By contrast, in spite of the heavy influence of this age group on the estimated regression coefficients, the death rate among 35–44-year-old smokers is seriously overestimated by the additive model. With the power model, the residuals are all quite small, indicative of the good fit, and none of them shows statistically significant deviations when referred to tables of the standard normal distribution.

An alternative method of examining the goodness-of-fit is *via* the introduction of regression variables representing the interaction of smoking and age. For this purpose we defined a single quantitative interaction term in coded age and exposure levels, namely $x_{jk} = (j - 3)(k - 1.5)$ for $j = 1, 2, \ldots, 5$ and $k = 1, 2$. The constants 3 and 1.5 were subtracted before multiplying in order that the interaction variable be not too highly correlated with the main effects for age and smoking (see §6.10, Volume 1). Table 4.9 shows the estimated regression coefficient $\hat{\gamma}$ of the interaction variable, its standard error, and the goodness-of-fit statistic G^2 for each of the three basic models. The deviances at the bottom of Table 4.7 show that introduction of the interaction term results in a marked improvement in fit for both additive and multiplicative models. Note that the estimated interaction term is positive in the additive model, indicating that the rate difference increases with age, and negative in the multiplicative model, indicating that the rate ratios decline. These

Table 4.8 Regression diagnostics and adjusted residuals for three statistical models fitted to the data on coronary deaths among British male doctors (Table 3.15)

Age range (years) (j)	Statistical model					
	Additive ($\rho = 1$)		Power ($\rho = 0.55$)		Multiplicative ($\rho = 0$)	
	Non-smokers ($k = 1$)	Smokers ($k = 2$)	Non-smokers ($k = 1$)	Smokers ($k = 2$)	Non-smokers ($k = 1$)	Smokers ($k = 2$)
Diagonal elements of hat matrix (h_{jk})						
35–44	0.98	0.93	0.67	0.86	0.25	0.81
45–54	0.31	0.77	0.42	0.85	0.29	0.88
55–64	0.19	0.82	0.28	0.85	0.38	0.91
65–74	0.18	0.83	0.22	0.84	0.36	0.91
75–84	0.22	0.78	0.24	0.79	0.34	0.87
Approximate change in β coefficient for smoking after deletion of each observation						
35–44	0.855	0.855	0.026	0.026	−0.060	−0.060
45–54	−0.058	−0.058	−0.021	−0.021	−0.072	−0.072
55–64	−0.020	−0.020	−0.010	−0.010	−0.012	−0.012
65–74	−0.008	−0.008	−0.010	−0.010	0.018	0.018
75–84	0.002	0.002	0.023	0.023	0.138	0.138
Adjusted residuals ($\tilde{r}_{jk}$)						
35–44	2.17	−2.17	0.25	−0.25	−2.14	2.14
45–54	−1.58	1.58	−0.36	0.36	−1.47	1.47
55–64	−1.47	1.47	−0.27	0.27	−0.17	0.17
65–74	−1.30	1.30	−0.44	0.44	0.29	−0.29
75–84	0.64	−0.64	1.38	−1.38	2.51	−2.51

Table 4.9 Fitting of a quantitative interaction variable in age × smoking to the data on coronary deaths among British male doctors (Table 3.15): regression coefficients ± standard errors for three statistical models

	Statistical model		
	Additive ($\rho = 1$)	Power ($\rho = 0.55$)	Multiplicative ($\rho = 0$)
Coefficient γ	0.732 ± 0.301	−0.024 ± 0.095	−0.309 ± 0.097
Goodness-of-fit G^2 on 3 degrees of freedom	2.16	2.08	1.55

features are of course already evident from the original data (Table 3.15). The fit to the power model was improved scarcely at all by the interaction terms. Thus, for these simple data, inclusion of interaction variables to measure lack of fit gives the same result as when lack of fit is evaluated *via* the power model.

4.5 Grouped data analyses of the Montana cohort with the multiplicative model

We now turn to a re-examination of the data on Montana smelter workers analysed in Chapter 3, in order to illustrate how the results of model fitting compare with the

techniques of standardization. Appendix V contains the data records from the Montana smelter workers study that were analysed earlier using standardization and related techniques (Tables 3.3, 3.11 and 3.16). There are $J = 16$ strata formed by the combination of four ten-year age groups and four calendar periods of variable length. The $K = 8$ 'exposure' categories are defined by the combination of two periods of first employment or date of hire (1 = pre-1925, 2 = post-1925) and four categories of duration of employment in work areas with high or medium exposure to arsenic (1 = <1 years, 2 = 1–4 years, 3 = 5–14 years, 4 = 15+ years). The coded levels for these four factors (AGE, YEAR, PERiod and EXPosure duration) appear as the first four columns (variables) in the data set. For 14 of the $4 \times 4 \times 2 \times 4 = 128$ combinations of these factors, no person-years of observation, and hence no deaths, occurred and no data record is included. These include the combinations with AGE = 1 (40–49 years), YEAR = 3 or 4 (1960–1977), and PER = 1 (pre-1925) and those with AGE = 2, YEAR = 4 and PER = 1. Individuals in these categories, being under 50 years of age in 1960, would have been aged 14 or less in 1925 and unlikely to have started work earlier. Likewise, there is no observation at AGE = 4 (60–79 years), YEAR = 1 (1938–1949), PER = 2 (post-1925) and EXP = 4 (15+ years).

(*a*) *Estimation of relative risk*

Our first goal is to reproduce as closely as possible the results obtained in the last chapter. Recall that relative risks of respiratory cancer for each duration of arsenic exposure were obtained separately for the pre-and post-1925 cohorts by three methods: (i) external standardization (Table 3.3); (ii) internal standardization (Table 3.11); and (iii) the Mantel–Haenszel procedure (Table 3.11). Maximum likelihood estimates of these same relative risks, using a multiplicative model with three binary exposure variables to represent the effect of each exposure category *versus* baseline (0–0.9 years heavy/medium arsenic exposure) and a varying number of stratum parameters α_j to represent the effects of age and calendar year, are shown in Table 4.10. Differences in the degrees of freedom for the goodness-of-fit statistics used with each subcohort are due to the fact that information on person-years was available for different combinations of age-year-exposure. For the pre-1925 cohort there were 13 age × year strata, and each of these had a data record for the full complement of four exposure categories. Thus, the total number of data records is $4 \times 13 = 52$ and the degrees of freedom are $52 - 13 - 3 = 36$. For the post-1925 cohort, two of the $4 \times 16 = 64$ possible exposure-age-year combinations were missing, and since there were $16 + 3 = 19$ parameters estimated, the degrees of freedom numbered $62 - 19 = 43$.

(*b*) *Testing for heterogeneity and trend in the relative risk with exposure duration*

The relative risk estimates and likelihood ratio (deviance) tests for heterogeneity and trend obtained *via* maximum likelihood fitting (Table 4.10) agree reasonably well with those based on Mantel–Haenszel methodology (Table 3.11). The Mantel–Haenszel style test statistics (3.24) and (3.25) are in fact efficient score tests based on the multiplicative model (4.2), as were the analogous statistics developed in Volume 1 for case-control data (Day & Byar, 1979). When interpreting the individual relative risk estimates by comparing each exposure group with baseline, it is important to

Table 4.10 Fitting of multiplicative models to grouped data from the Montana smelter workers study: internal estimation of baseline rates

Variable fitted	Relative risk (exponentiated regression coefficient) and standardized regression coefficient (in parentheses)			
	Employed prior to 1925	Employed 1925 or after	Combined cohort	
			Four levels of exposure	Two levels of exposure
Exposure duration (years)				
Under 1	1.0	1.0	1.0	1.0
1–4	2.43 (3.20)	2.10 (3.88)	2.21 (5.04)	
5–14	1.96 (2.15)	1.67 (1.84)	1.77 (2.77)	2.19 (6.45)
15+	3.12 (5.10)	1.75 (1.52)	2.58 (5.25)	
Pre-1925 employment	—	—	1.62 (3.23)	1.66 (3.46)
Deviance (G^2)	32.9	56.0	96.8	99.2
Degrees of freedom	36	43	94	96
Tests of significance of exposure based on G^2				
Global	$\chi^2_3 = 28.3$	$\chi^2_3 = 15.8$	$\chi^2_3 = 42.4$	$\chi^2_1 = 39.7$
Trend	$\chi^2_1 = 24.7$	$\chi^2_1 = 8.9$	$\chi^2_1 = 32.7$	

remember that they utilize information from all the exposure categories, and not just the two in question. For example, the estimated risk ratio of $\hat{\psi}_2 = 2.43$ comparing rates in the 1–4-year exposure duration category to those in the under-1-year category uses some information from the comparisons of the 1–4 *versus* 5–9 and under 1 *versus* 5–9 categories, and so on. Were we to estimate the relative risks for pairwise comparisons of exposure categories using only the data for each pair, whether by Mantel–Haenszel or maximum likelihood, the resulting estimates would fail to be consistent with each other. The product of estimated relative risks for under 1 *versus* 1–4 and 1–4 *versus* 5–9 years would not necessarily equal the relative risk for under 1 *versus* 5–9. The same phenomenon was noted also for case-control studies (§§4.5 and 5.5, Volume 1). Consistency is achieved only by building it into the fitted model.

(c) *Evaluating the goodness-of-fit of the multiplicative model*

An evaluation of the goodness-of-fit of the multiplicative model was made by examination of residuals and the addition of interaction terms to the model equation. While the goodness-of-fit statistic for the pre-1925 cohort is slightly less than its degrees of freedom, indicating that the model fits reasonably well overall, that for the post-1925 cohort is larger ($G^2 = 56.0$, degrees of freedom $= 43$, $p = 0.09$). The corresponding chi-square statistic is $\chi^2_{43} = 54.9$. However, examination of the observed and fitted numbers of deaths for the 62 age-year-exposure cells for this cohort reveals no particular pattern to the lack of fit. The greatest contribution to chi-square is from the 15+ year exposure category for ages 50–59 and years 1950–1959 where two respiratory cancer deaths were observed *versus* 0.24 expected from the multiplicative model. Elimination of this one cell would markedly improve the fit. The example also serves as

Table 4.11 Evaluating the goodness-of-fit of the multiplicative model of Table 4.10; deviance test statistics for interaction effects

Interaction effect	Degrees of freedom	Employed prior to 1925	Employed 1925 or after	Combined cohort
Year × exposure				
Qualitative	9	8.2	18.0	15.3
Quantitative (linear × linear)	1	2.7	2.5	6.5
Age × exposure				
Qualitative	9	9.1	14.0	13.2
Quantitative (linear × linear)	1	3.3	3.7	3.5

a reminder that the usual asymptotic approximations for χ^2 and G^2 statistics may not apply when the data are sparse and expected values for some cells are small (McCullagh, 1986).

In order to look more systematically for possible trends in the relative risks with age and year, we examined a number of additional models with both qualitative and quantitative interaction terms. The results, summarized in Table 4.11, do not suggest that the relative risks estimated for different exposure durations change systematically with either age or year in the pre-1925 cohort. However, even in the absence of a definite trend, there is considerable variation in the exposure effects from one calendar period to another for the post-1925 cohort. Some caution needs to be exercised, therefore, in interpreting the relative risks shown in Table 4.10 for the latter cohort.

Table 4.12 lists the deviances for several models that we fitted to the full set of cohort data in the process of obtaining the results shown in the right-hand columns of

Table 4.12 Goodness-of-fit (deviance) statistics for a series of models fitted to the data on Montana smelter workers: internal estimation of baseline rates

Model number	Terms included in the model[a]	Degrees of freedom	Deviance
1	—	98	155.4
2	PER	97	138.9
3	EXP	95	107.0
4	PER + EXP	94	96.8
5	PER + EXP0 + PER . EXP	91	94.6
6	PER + EXP1	96	99.2
7	PER + EXP1 + PER . EXP1	95	97.5
8	PER + EXP1 + AGE . EXP1	93	90.0
9	PER + EXP1 + YEAR . EXP1	93	95.2

[a] In addition to 16 terms for stratum (age and year) effects. The variables are coded as follows: PER, period of employment (pre- *versus* post-1925); EXP, four-level factor for duration of exposure; EXP1, binary indicator of one or more years of heavy/medium arsenic exposure

Table 4.10. (Models 4 and 6 in Table 4.12 correspond to columns 3 and 4 of Table 4.10.) The first four lines of Table 4.12 show that both period of first employment and exposure to heavy-medium arsenic had marked and relatively independent effects on risk. The addition of period × exposure interaction terms (model 5) does not significantly improve the fit. Relative risks for the three exposure duration levels are estimated by model 5 to be 2.43, 1.99 and 3.22 for those first employed before 1925 and 2.03, 1.63 and 1.62 for those employed afterwards, which results compare well to those obtained when the two subcohorts are analysed separately (columns 1 and 2 of Table 4.10).

An important advantage of model fitting is the flexibility it offers for looking at the same data in a number of different ways. Examination of the relative risk estimates in Table 4.10 suggests that they do not change much either with increasing duration of exposure or with period of first employment. In order to study the issue further, we constructed a new binary exposure variable EXP1 to indicate whether or not a full year of heavy/medium arsenic exposure had yet been experienced, and fitted several additional models to the complete set of cohort data. The most interesting aspect of Table 4.12 is the comparison of models 5 and 6. Constraining the relative risk estimates for arsenic exposure to be constant regardless of period or duration of exposure leads to nearly as good a summary of the data as allowing them to vary ($99.2 - 94.6 = 4.6$, 5 degrees of freedom, $p = 0.47$). The estimated effect of exposure for a year or more to heavy/medium levels of arsenic is to increase the subsequent respiratory cancer death rate by a factor of 2.2. There is little evidence that this estimate of arsenic effect changes with additional exposure or according to the date of hire. The improved fit from model 8 suggests, however, that it may depend on age ($99.2 - 90.0 = 9.2$, 3 degrees of freedom, $p = 0.03$), the estimated relative risks being 1.68, 3.07, 2.50 and 1.10 for the four age groups. Using EXP1 rather than EXP gives less evidence for an interaction with calendar year; the separately estimated relative risks for the four decades are 2.57, 3.23, 1.92 and 1.76.

In order to determine whether one or two data records might have had an undue influence on the fit, we computed the 'hat' matrix elements and approximate changes in regression coefficients for model 6 of Table 4.12. (This model is also shown in the last column of Table 4.10.) GLIM was used to carry out the calculations of h and $\Delta\hat{\beta}$ using equations 4.14 and 4.15. As expected, the records with the largest effects on the overall fit were generally those with the largest person-years of observation: record 17 with seven lung cancer cases and over 12 000 person-years gave $h = 0.617$; record 21 with one case and 7151 person-years gave $h = 0.634$; and record 49 with 89 cases and 8495 person-years gave $h = 0.590$. The total value of h summed over all 114 records is 18, the number of parameters being estimated.

Other data records had the largest influence on the estimated effect of arsenic exposure as evaluated by the change in the coefficient of EXP1. The maximum change occurred with record 56 (9 cases observed *versus* 4.00 expected), the deletion of which would reduce the relative risk associated with heavy/moderate arsenic exposure by a factor of approximately $\exp(-0.086)$, i.e., from 2.19 to 2.01. None of these results suggests any serious instability in the fitted model.

4.6 Incorporating external standard rates into the multiplicative model

Up to now we have considered that the stratum-specific parameters α_j in the model equation (4.10), which represent the log death rates for unexposed ($x_{jk} = 0$) individuals in that stratum, were unknown 'nuisance' parameters to be estimated internally from the study data. The spirit of this approach is similar to that discussed in §§3.5 and 3.6. It avoids the problems caused by the noncomparability of external standard rates, namely, that relative risk estimates for different exposure groups will fail to summarize adequately the stratum-specific rate ratios.

While this ability of multivariate modelling to accommodate the internal estimation of baseline rates is desirable, incorportion of external standard rates into the analysis may be advantageous in some circumstances. Suppose the baseline rates are specified up to a scale factor θ, say $\lambda_j = \theta\lambda_j^*$ where the λ_j^* are known from vital statistics or other sources. The model equation analogous to (4.10) is

$$\log \lambda_{jk} = \alpha_j^* + \mu + \mathbf{x}_{jk}\boldsymbol{\beta}, \tag{4.18}$$

where $\mu = \log(\theta)$ is a parameter (the grand mean) which represents the log SMR for the unexposed ($x_{jk} = 0$), and $\alpha_j^* = \log \lambda_j^*$. If follows that the mean values $E(d_{jk})$ for the number of deaths in the (j, k) cell satisfy

$$\log E(d_{jk}) = \log(n_{jk}\lambda_j^*) + \mu + \mathbf{x}_{jk}\boldsymbol{\beta},$$

so that now the log expected standard deaths are declared as the OFFSET in a GLIM analysis, rather than the log person-years (compare equation 4.11).

One advantage of (4.18) is that it provides in the parameter μ an overall measure of how the baseline cohort rates compare with those for the general population. Also, since the number of parameters to be estimated from the data is reduced considerably in comparison to (4.11), there could theoretically be an improvement in the efficiency of estimation of the β parameters of most interest. However, this improvement is not likely to be great for many practical problems (see Example 4.7). Perhaps more important is the fact that when the x variables depend only on exposure (k) and not on stratum (j), the likelihood for the model (4.18) is a function of the totals $O_k = \sum_j d_{jk}$ and $E_k^* = \sum_j n_{jk}\lambda_j^*$ of observed and expected deaths in each of the K exposure categories. (In fact, if a separate parameter β_k is attached to each exposure category, the $\mathrm{SMR}_k = O_k/E_k^*$ are maximum likelihood estimates.) This permits a much more economical presentation of the basic data needed for the regression analysis than is true for the models considered in the preceding section. For tests of goodness-of-fit, however, the full set of data records for all $J \times K$ cells are needed.

Since the β parameters describe how the log SMR varies as a function of the exposures, (4.18) extends the method of indirect standardization into the domain of multivariate regression analysis. If $\boldsymbol{\beta}$ indexes K different exposure classes, the efficient score statistic of the hypothesis $\boldsymbol{\beta} = \mathbf{0}$ developed from this model corresponds to the statistic (3.11) previously proposed for testing heterogeneity of risk. Likewise, for a single quantitative regression variable the score test of $\boldsymbol{\beta} = \mathbf{0}$ is identical with the trend test (3.12). Finally, the maximum likelihood estimate of μ in the model where $\boldsymbol{\beta} = \mathbf{0}$ is

precisely $\hat{\mu} = \log(O_+/E_+^*) = \log(\text{SMR})$, where O_+ and E_+^* denote totals of observed and expected values. These results provide the essential link between the elementary methods of cohort analysis considered in Chapter 3 and those based on the multiplicative model.

The drawbacks of indirect standardization noted earlier of course apply also to an uncritical application of the regression model. However, the process of model fitting encourages the investigator to evaluate the assumptions of proportionality that are essential in order that the estimated β parameters have the intended interpretation. The usual goodness-of-fit machinery may be applied to validate these assumptions. Additional terms may be incorporated in the model to account for confounding of the SMR/exposure relationship by age, year or other stratification factors. The estimates of the exposure effects as expressed in $\hat{\boldsymbol{\beta}}$ will then start to approximate those obtained with the model (4.10), wherein the baseline rates are estimated internally. See §4.8 for an example.

Example 4.3

To illustrate the process of multivariate modelling using external standard rates, we return to the problem of estimating relative risks of respiratory cancer associated with duration of heavy/medium arsenic exposure in the Montana cohort. The basic data needed to fit the models consist of just eight records containing observed (O_k) and expected (E_k^*) numbers of deaths by period of employment and exposure duration (Table 3.3).

Table 4.13 summarizes the results of fitting the same models as in Table 4.10, but where the baseline rates are obtained from Table 3.2 rather than estimated internally. There is good agreement between the two analyses as far as the arsenic effects are concerned, but the pre- *versus* post-1925 period effect is overestimated when the comparison is made using the external standard rates. Goodness-of-fit using external

Table 4.13 Fitting of multiplicative models to grouped data from the Montana smelter workers study: external baseline rates

Regression variable	Relative risk (exponentiated regression coefficient) and standardized regression coefficient (in parentheses)			
	Employed prior to 1925	Employed 1925 or after	Combined cohort	
			Four levels of exposure	Two levels of exposure
Constant (SMR)	2.38 (6.18)	1.35 (2.99)	1.28 (2.65)	1.26 (2.55)
Duration heavy/medium arsenic exposure (years)				
Under 1	1.0	1.0	1.0	1.0
1–4	2.43 (3.17)	2.05 (3.78)	2.16 (4.89)	
5–14	1.99 (2.21)	1.63 (1.76)	1.76 (2.73)	2.16 (6.36)
15+	3.22 (5.28)	1.62 (1.31)	2.58 (5.28)	
Pre-1925 employment	—	—	2.03 (5.59)	2.09 (5.99)
Deviance (G^2)	39.0	70.4	112.3	114.7
Degrees of freedom	48	58	109	111
Tests of significance of exposure based on G^2				
Global	$\chi_3^2 = 29.6$	$\chi_3^2 = 14.4$	$\chi_3^2 = 41.0$	$\chi_1^2 = 38.6$
Trend	$\chi_1^2 = 26.1$	$\chi_1^2 = 7.6$	$\chi_1^2 = 32.4$	

standard rates appears no worse than when the rates are estimated. Note that we have considered the fit of the model to the original data from Appendix VI rather than to the summary data in Table 3.3 in order to be able to evaluate goodness-of-fit.

The relative risk estimates shown separately for the two employment periods in Table 4.13 are identical to those given in Table 3.3, and the coefficients $\exp(\hat{\mu}) = 2.38$ for the pre-1925 cohort or $\exp(\hat{\mu}) = 1.35$ for the post-1925 cohort also agree with the SMRs of 238% and 135% found earlier for the baseline exposure duration category (under 1 year). This is a numerical confirmation of the fact that the maximum likelihood estimates of parameters in these simple qualitative models are log (SMR)s, or differences between log (SMR)s.

4.7 Proportional mortality analyses

Regression analyses similar to those already considered for grouped cohort data with person-years denominators can also be carried out using information only for persons who have died. As mentioned in §3.7, the data are best considered as arising from a case-control study in which the persons who die from the cause of interest are regarded as the 'cases', while those who die of other causes (or some subset thereof) are the 'controls'. They are classified into precisely the same J strata and K exposure classes as are the cases and person-years in the corresponding cohort analysis. The observations in stratum j and exposure class k consist of the number d_{jk} of deaths or cases, the total t_{jk} of cases and controls (all deaths) and the associated covariables $\mathbf{x}_{jk}$.

(a) *Derivation of the logistic regression model*

As usual we denote the death rate from the cause of interest in the (j, k) cell by λ_{jk}. We denote the death rate from the other causes by ν_{jk} so that the total death rate is given by $\lambda_{jk} + \nu_{jk}$. Let us suppose that each of these satisfies the multiplicative model (4.10), say

$$\log \lambda_{jk} = \alpha_j + \mathbf{x}_{jk}\boldsymbol{\beta}$$
$$\log \nu_{jk} = \gamma_j + \mathbf{x}_{jk}\boldsymbol{\delta}. \tag{4.19}$$

It follows that the conditional probability p_{jk} that a death in the (j, k) cell is from the cause of interest, given that one occurred at all, is given by

$$p_{jk} = \frac{\lambda_{jk}}{\lambda_{jk} + \nu_{jk}} = \frac{\exp\{(\alpha_j - \gamma_j) + \mathbf{x}_{jk}(\boldsymbol{\beta} - \boldsymbol{\delta})\}}{\exp\{(\alpha_j - \gamma_j) + \mathbf{x}_{jk}(\boldsymbol{\beta} - \boldsymbol{\delta})\} + 1}.$$

In other words, the probability that a death is from the specific cause satisfies the linear logistic model

$$\operatorname{logit} p_{jk} = \log \frac{p_{jk}}{1 - p_{jk}} = (\alpha_j - \gamma_j) + \mathbf{x}_{jk}(\boldsymbol{\beta} - \boldsymbol{\delta}). \tag{4.20}$$

Furthermore, if the exposures have no effect on the rate of death from the other causes ($\boldsymbol{\delta} = \mathbf{0}$), the regression parameters of the covariables $\mathbf{x}_{jk}$ estimated from this linear logistic relationship correspond precisely to the log relative risks of principal interest. This provides a formal confirmation of the well-known fact that proportional mortality analyses are valid only if the controls are selected from among deaths due to causes

Table 4.14 Fitting of multiplicative models to grouped data from the Montana smelter workers study: proportional mortality analysis with internal control

Variable fitted	Relative risk (exponentiated regression coefficient) and standardized regression coefficient (in parentheses)			
	Employed prior to 1925	Employed 1925 or after	Combined cohort	
			Four levels of exposure	Two levels of exposure
Exposure duration (years)				
Under 1	1.0	1.0	1.0	1.0
1–4	2.32 (2.69)	2.02 (3.40)	2.21 (4.37)	
5–14	1.98 (1.98)	1.59 (1.56)	1.74 (2.49)	2.07 (5.54)
15+	2.82 (4.20)	1.60 (1.21)	2.32 (4.22)	
Pre-1925 employment	—	—	1.54 (2.66)	1.56 (2.81)
Deviance (G^2)	47.0	48.6	104.8	106.0
Degrees of freedom	35	41	91	93
Tests of significance of exposure based on G^2				
Global	$\chi^2_3 = 20.6$	$\chi^2_3 = 12.2$	$\chi^2_3 = 31.0$	$\chi^2_1 = 29.8$
Trend	$\chi^2_1 = 18.9$	$\chi^2_1 = 6.4$	$\chi^2_1 = 23.5$	

that have no relation to the exposures. Prentice and Breslow (1978) make the same observation in deriving the analogous relationship for continuous data.

In order to carry out the proportional mortality analysis, we treat the d_{jk} as independent binomial random variables with denominators t_{jk} and probabilities p_{jk} of 'being a case' that satisfy the linear logistic model (4.20). Techniques of maximum likelihood estimation are applied exactly as described in Chapter 6 of Volume 1. Provided that the other causes of death are unrelated to the exposures, the regression coefficients may be interpreted as log relative risks in the usual fashion.

Example 4.4

The data in Appendix V include the total numbers of deaths observed in each of the 114 categories defined by the cross-classification in the Montana smelter workers study. These were analysed using the logistic regression model (4.20) with covariables $\mathbf{x}_{jk}$ defined just as in the earlier cohort analyses to represent the effects of period of hire and duration of moderate to heavy arsenic exposure. Table 4.14 presents the results in the same format as for the parallel cohort analyses (Table 4.10). The significance of the estimated exposure effects is somewhat reduced in comparison, as might be expected since more restricted data are being used. The deviances measuring the goodness-of-fit of the models to the proportional data are considerably higher. Note that three degrees of freedom have been lost in comparison with Table 4.10, due to the fact that there was no death at all ($t_{jk} = 0$) in three cells. Nevertheless, the estimated regression coefficients for the proportional mortality analysis are quite comparable to those for the full cohort analysis. There is a slight reduction in the estimated effects for period of hire and for 15 or more years of arsenic exposure, indicating that these two factors may possibly have increased mortality rates from causes other than respiratory cancer.

(*b*) *Incorporating standard rates into the proportional mortality analysis*

Suppose now that external standard rates λ_j^* and ν_j^* are available for deaths due to specific and nonspecific causes in stratum j. We continue to rely on the basic multiplicative model (4.19), except that the unknown log background rates α_j and γ_j

are replaced by $\alpha + \log \lambda_j^*$ and $\gamma + \log \nu_j^*$, respectively. Defining $p_j^* = \lambda_j^*/(\lambda_j^* + \nu_j^*)$ to be the standard proportions of deaths due to the cause of interest in the jth stratum, it follows that the probability that a death is due to that cause may be written

$$\text{logit}\, p_{jk} = \text{logit}\, p_j^* + (\alpha - \gamma) + \mathbf{x}_{jk}(\boldsymbol{\beta} - \boldsymbol{\delta}). \tag{4.21}$$

The probabilities of 'being a case' continue to satisfy the linear logistic model (4.20). Now, however, the known variable logit p_j^* 'offsets' the model equation, and there is a constant term with coefficient $(\alpha - \gamma)$. This coefficient may be interpreted as the logarithm of the standardized relative mortality ratio (SRMR) for unexposed members of the cohort ($\mathbf{x} = \mathbf{0}$), where the SRMR is defined as the ratio of SMRs for specific *versus* nonspecific causes (Breslow & Day, 1975). The proportional mortality ratio as usually defined, namely the ratio of the number of deaths observed to those 'expected' on the basis of the stratum-specific proportions p_j^*, is of lesser interest for reasons discussed in §3.7.

Example 4.5

Table 4.15 presents the standard proportions p_j^* for the 16 age × year strata used with the Montana smelter workers data. These were obtained by dividing the standard death rates from respiratory cancer (Table 3.2) by the corresponding standard death rates for all causes. The logistic transform of these standard proportions was used as an offset in a logistic regression analysis based on equation (4.21).

Table 4.16 presents the results in what has now become a standard format. Just as observed earlier for the full cohort analysis (Tables 4.10 and 4.13), the relative risk estimated for pre- *versus* post-1925 employment is greater when the standardized proportions are used as a basis of comparison than when these same proportions are estimated internally. This suggests that the association between the SMR (or SRMR) and period of employment is confounded by one or more of the stratification factors, an interpretation that is confirmed by more detailed analyses of data from the Montana cohort reported below. Otherwise, the agreement between the two types of proportional mortality analyses is quite good. The difference between the constant terms for the full cohort analysis (Table 4.13) and the proportional analysis (Table 4.16) suggests that the γ coefficient in equation (4.21) is nonzero. This implies simply that the SMR for nonrespiratory cancer deaths among cohort members with zero covariates is different from unity.

4.8 Further grouped data analyses of the Montana cohort

The preceding illustrative analyses of the Montana smelter workers study are limited in scope by the requirement that they be based on the relatively small data set presented in Appendix V. More realistic analyses were undertaken also by fitting multiplicative models to a more elaborate set of grouped data (Breslow, 1985a; Breslow & Day, 1985). Respiratory cancer deaths and person-years of exposure were

Table 4.15 Standard proportions of deaths due to respiratory cancer: US white males

Age group (years)	Calendar year			
	1938–1949	1950–1959	1960–1969	1970–1977
40–49	0.021515	0.038246	0.052288	0.070208
50–59	0.028478	0.055765	0.074081	0.095478
60–69	0.021247	0.047646	0.072328	0.095159
70–79	0.009894	0.024390	0.041900	0.064688

Table 4.16 Fitting of multiplicative models to grouped data from the Montana smelter workers study: proportional mortality analysis with external control

Variable fitted	Relative risk (exponentiated regression coefficient) and standardized regression coefficient (in parentheses)			
	Employed prior to 1925	Employed 1925 or after	Combined cohort	
			Four levels of exposure	Two levels of exposure
Constant (SRMR)	2.02 (4.77)	1.13 (1.16)	2.22 (6.54)	2.26 (6.83)
Exposure duration (years)				
Under 1	1.0	1.0	1.0	1.0
1–4	2.26 (2.64)	1.90 (3.14)	2.02 (4.11)	
5–14	2.15 (2.25)	1.49 (1.36)	1.72 (2.45)	2.03 (5.42)
15+	2.89 (4.37)	1.54 (1.11)	2.35 (4.31)	
Pre-1925 employment	—	—	2.07 (5.33)	2.13 (5.74)
Deviance (G)	55.9	65.6	123.8	125.2
Degrees of freedom	47	56	106	108
Tests of significance of exposure based on G				
Global	$\chi^2_3 = 21.8$	$\chi^2_3 = 10.3$	$\chi^2_3 = 29.8$	$\chi^2_1 = 28.4$
Trend	$\chi^2_1 = 19.6$	$\chi^2_1 = 5.3$	$\chi^2_1 = 23.7$	

classified in six dimensions: (i) age in four ten-year intervals; (ii) calendar year in four intervals; (iii) date of first employment (pre- *versus* post-1925); (iv) birthplace (US- *versus* foreign-born); (v) number of years worked in moderate arsenic areas (<1, 1–4, 5–14, 15+) and (vi) number of years worked in heavy arsenic areas (<1, 1–4, 5+). Of the $4 \times 4 \times 2 \times 2 \times 4 \times 3 = 768$ possible cells in this six-dimensional table, only 478 actually contained any person-years observation. The results obtained in this section will serve as a useful point of reference for those based on more complicated methods of analysis of continuous data that are considered in the next chapter.

(*a*) *Preliminary analyses*

Table 4.17 presents respiratory cancer SMRs according to a large number of possible risk variables, including several not mentioned above, and without regard to possible confounding effects. The time-dependent exposure variables were lagged two years in an attempt to estimate the exposure status at the time of disease onset, rather than at the time of death, and thus to avoid some of the healthy worker selection problem. Tests of significance were based on the heterogeneity statistic (3.11), or the trend test (3.12), as appropriate.

From this preliminary analysis, we conclude that period of first employment, birthplace, years since first employed and level of arsenic exposure may each have some effect on the age-specific rates. We also note a sharp decline in the SMR with calendar year, indicating that the respiratory cancer rates for the cohort as a whole have not increased in constant proportion with those for the general population, although they have remained consistently higher (see Example 2.4). This serves as a warning of a possible lack of comparability of the SMRs for the various exposure

Table 4.17 Variations in respiratory cancer SMRs among Montana smelter workers[a]

Factor analysed	Level	Number of deaths	SMR(×100)[b]	Test of significance[c]
Period of first employment	1885–1924	115	362	$\chi^2_1 = 39.5$
	1925–1955	161	164	$(p < 0.0001)$
Age at hire (years)	<24	69	255	$\chi^2_1 = 5.2$
	25–34	116	222	$(p = 0.07)$
	35+	91	184	
Birthplace	US	198	180	$\chi^2_1 = 28.5$
	Foreign	80	381	$(p < 0.0001)$
Time since first employed[d] (years)	1–14	101	165	$\chi^2_2 = 24.0$
	15–29	59	185	$(p < 0.0001)$
	30+	116	315	
Time since last employed[d] (years)	None	110	230	$\chi^2_2 = 3.2$
	0–9	84	227	$(p = 0.20)$
	10+	82	181	
Arsenic exposure[d]	Light only	153	160	$\chi^2_2 = 44.4$
	Moderate[e]	91	339	$(p < 0.0001)$
	Heavy[e]	32	434	
Age at follow-up (years)	40–49	21	166	$\chi^2_3 = 2.5$
	50–59	80	199	$(p = 0.48)$
	60–69	117	228	
	70–79	58	223	
Year at follow-up	1938–1949	34	403	$\chi^2_3 = 28.4$
	1950–1959	65	294	$(p < 0.0001)$
	1960–1969	94	211	
	1970–1977	83	151	

[a] From Breslow (1985a)
[b] Calculated with reference to US mortality rates for white males by age and calendar year
[c] Test for homogeneity of SMRs among categories shown based on equations (3.11) or (3.12)
[d] Time-dependent exposure variable lagged two years
[e] Worked in moderate or heavy arsenic exposure area for at least one year

classes due to confounding with calendar year. Additional confounding may result from the high correlation between certain exposure variables. For example, due to the fact that follow-up started only in 1938, virtually everyone employed before 1925 contributed person-years only to the last two categories of years since first employment.

Table 4.18 presents an analysis of variance of the log SMRs based on the model equation (4.18) and various indicator regression variables. This shows clearly that the effects of duration of employment are easily explained by the correlation with period of first employment, whereas those for arsenic exposure are not. Selection of the variables for the final analysis was based on such considerations.

(*b*) *Regression analyses*

Table 4.19 presents further multiple regression analyses based on equations (4.18) (first two columns) and (4.11) (last column). When calendar year is included in the

Table 4.18 Analysis of variance based on a multiplicative model for SMRs: respiratory cancer deaths in Montana smelter workers[a]

Source of variation[b]	Degrees of freedom	Chi-square
PERIOD of hire and TIME since first employment	3	41.0
PERIOD alone	1	39.5
TIME after PERIOD	2	1.5[c]
TIME alone	2	24.0
PERIOD after TIME	1	17.0
PERIOD and ARSENIC level	3	77.6
PERIOD alone	1	39.5
ARSENIC after PERIOD	2	38.1
ARSENIC alone	2	44.4
PERIOD after ARSENIC	1	33.2

[a] From Breslow (1985a)
[b] See Table 4.17 for definition of factor levels
[c] Not statistically significant; all others have $p < 0.0001$

SMR analysis, the regression coefficient for period of hire is much closer to that obtained when baseline rates are estimated internally. This confirms that part of the difference between the SMRs for those hired before and after 1925 is due to the confounding effects of calendar year on the ratios of cohort to standard death rates. Appropriate adjustment is made either by including calendar year as a covariable in

Table 4.19 Regression coefficients ± standard errors in the multiplicative model: two methods of analysis of grouped data from the Montana smelter workers study[a]

Regression variable	Method of analysis		
	External standard rates (SMR analysis)		Internal estimation of baseline rates by age and year
	Without calendar year effects	With calendar year effects	
Constant (α)	0.256 ± 0.092	0.581 ± 0.219	—
Hired before 1925	0.564 ± 0.133	0.441 ± 0.143	0.444 ± 0.151
Foreign born	0.492 ± 0.142	0.407 ± 0.147	0.445 ± 0.153
Heavy arsenic			
1–4 years	0.170 ± 0.310	0.199 ± 0.303	0.193 ± 0.305
5+ years	1.067 ± 0.230	1.076 ± 0.230	1.069 ± 0.230
Moderate arsenic			
1–4 years	0.587 ± 0.166	0.604 ± 0.166	0.600 ± 0.166
5–14 years	0.253 ± 0.242	0.262 ± 0.242	0.259 ± 0.242
15+ years	0.678 ± 0.204	0.683 ± 0.205	0.689 ± 0.206
Calendar period			
1950–1959		−0.075 ± 0.216	
1960–1969		−0.235 ± 0.215	
1970–1977		−0.480 ± 0.228	

[a] From Breslow and Day (1985)

the SMR analysis or else, and what is almost the same, by conducting a parallel internally controlled analysis in which background rates are estimated from the data. If age, year and age × year interactions are all included as covariables in the SMR analysis, the externally and internally controlled analyses yield identical results as far as the β coefficients are concerned. The difference in the coefficients for foreign-born between the second and third columns suggests some possible residual confounding by age.

4.9 More general models of relative risk

One possible drawback to the multiplicative model (4.10), at least when applied with quantitative exposure variables, is that it leads to relative risk functions that increase exponentially with increasing exposure: $\mathrm{RR}(x) = \exp(x\beta)$. Apparently, some risks do increase this fast. For example, our analyses of the Ille-et-Vilaine case-control study in Volume 1 showed that alcohol had such an effect on the risk of oesophagael cancer. This example is atypical, however, and in most epidemiological studies the rate of increase would be less dramatic (see Chapter 6). In Ille-et-Vilaine, the relative risk of oesophageal cancer was approximately proportional to the square root of the daily dose of tobacco.

(*a*) *Transformations of dose*

Many of the quantitative dose-response relations actually observed in cancer epidemiology approximate a power relationship of the form

$$\mathrm{RR}(x) = (x + x_0)^{\beta}. \tag{4.22}$$

Here $x_0 > 0$ is a small 'background' exposure level introduced to account for the spontaneous incidence of cancer among the unexposed. This relative risk function may be approximated by first transforming the dose to $z = \log(x + x_0)$ and then fitting the multiplicative model (4.10) in the form

$$\log \lambda_{jk} = \alpha_j + z_k\beta = \alpha_j + \{\log(x_k + x_0)\}\beta. \tag{4.23}$$

The choice $x_0 = 1$ is not uncommon as a 'starter' dose since it yields the usual $\mathrm{RR}(x) = 1$ at the baseline level $x = 0$. x_0 may also be treated as an unknown parameter and the best fitting value found by trial and error or some other more systematic technique. However, the model is then no longer a log-linear one, and determination of the variances and covariances of the parameter estimates may be seriously complicated. There is a high degree of correlation between the estimates of x_0 and β, as might be expected from the fact that the slope of the relative risk function (4.22) at $x = 0$ is given by $\mathrm{RR}'(0) = \beta x_0^{1-\beta}$. Since small variations in x_0 often have little effect on the overall goodness-of-fit, it is usually adequate simply to select a nominal background dose *a priori* and to proceed assuming that x_0 is fixed.

(*b*) *Additive relative risk model*

Certain formulations of multistage theory and other more general considerations lead to relative risk functions that are linear or quadratic in measured exposures, for example $\mathrm{RR}(x) = 1 + \beta x$ or $\mathrm{RR}(x) = 1 + \beta x + \gamma x^2$ (Berry, 1980; Thomas, D.C., 1981). These are special cases of a general class of models of the form

$$\lambda_{jk} = \exp(\alpha_j)\{1 + \mathbf{x}_{jk}\boldsymbol{\beta}\} \tag{4.24}$$

which we shall call the *additive relative risk model.* One drawback of these is that the range of the $\boldsymbol{\beta}$ parameters is necessarily restricted by the requirement that $\mathbf{x}_{jk}\boldsymbol{\beta} > -1$ for all values of $\mathbf{x}_{jk}$, since negative relative risks would otherwise result. This suggests that, wherever possible, the regression variables $\mathbf{x}_{jk}$ be coded so that they have positive coefficients. As usually happens for models in which there is a range restriction on the parameters, the log-likelihood function is skewed and not at all like the quadratic, symmetric log-likelihood of the approximating normal distribution. Estimates of the parameters may be unstable, and standard errors that are determined from the usual likelihood calculations may be unhelpful in assessing the degree of uncertainty. (This contrasts with additive models for absolute risk, where t statistics perform reasonably well. See Example 4.2.) Substantial differences may exist in practice between the observed and expected information measures, and score tests based on the former may give seriously misleading answers (Storer *et al.*, 1983). Irregularities in the likelihood surface may frustrate the search for maximum likelihood estimates. The usual iterative procedures can diverge unless starting values in the immediate vicinity of the maximum likelihood are available.

In view of these complications, we do not recommend the additive relative risk model for routine applications. It often suffices to transform the exposure variables and to approximate the additive relative risk model by a multiplicative model in the transformed variables. Nevertheless, one sometimes finds that the extra work involved in fitting the model (4.24) results in a substantially better fit to the data or is necessary in order that the regression coefficients have precisely the desired interpretation.

(*c*) *Fitting general models to Poisson rates*

The additive relative risk model (4.24) is a generalized linear model that involves a composite link function (Thompson & Baker, 1981): two separate linear functions (linear predictors) of the explanatory variables are related to the mean values $E(d_{jk}) = n_{jk}\lambda_{jk}$. Since it is a nonlinear regression model for Poisson rates, however, it still may be fitted using GLIM or other programs that facilitate iterated reweighted least-squares analyses (Frome, 1983). However, the implementation is more involved than for the multiplicative (4.16) or power (4.15) models considered earlier.

In their most general form Poisson regression models may be written

$$\lambda_{jk} = g(\mathbf{x}^*_{jk}; \boldsymbol{\beta}^*), \tag{4.25}$$

where the asterisks on $\boldsymbol{\beta}^*$ and $\mathbf{x}^*_{jk}$ indicate that these are the expanded vectors of length $J + p$ that involve the J stratum indicators and associated coefficients α_j in addition to the exposure variables. Maximum likelihood fitting of such models can be programmed

as a series of weighted least-squares analyses involving dependent variables

$$y_{jk} = d_{jk} - n_{jk}\hat{\lambda}_{jk},$$

weights

$$w_{jk} = (n_{jk}\hat{\lambda}_{jk})^{-1}$$

and independent variables

$$z_{jk} = n_{jk}\frac{\partial g(\mathbf{x}^*_{jk}; \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*}$$

that are recalculated at each stage of iteration using fitted rates

$$\hat{\lambda}_{jk} = g(\mathbf{x}^*_{jk}; \hat{\boldsymbol{\beta}}^*)$$

based on the current parameter estimates $\hat{\boldsymbol{\beta}}^*$. The change in the estimated coefficient going from one iteration to the next is given by

$$\Delta\hat{\boldsymbol{\beta}}^* = (\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^\mathsf{T}\mathbf{W}y,$$

where $\mathbf{Z}$ is a matrix with rows $\mathbf{z}_{jk}$ and $\mathbf{W}$ is a diagonal matrix with diagonal elements w_{jk}.

Programming the likelihood calculations in this fashion leads to regression diagnostics that help evaluate the goodness-of-fit and stability of the model just as we saw earlier for the multiplicative model and the power family (4.16). The diagonal terms h_{jk} of the 'hat' matrix obtained at convergence of the iterative procedure,

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{Z}(\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^\mathsf{T}\mathbf{W}^{1/2}, \tag{4.26}$$

provide information about the general influence of the data in cell (j, k) on the fit. The specific changes in the estimated regression coefficients occasioned by deletion of those data are approximated by the vector

$$(\Delta\hat{\boldsymbol{\beta}}^*)_{-jk} \simeq -(\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{Z})^{-1}z_{jk}y_{jk}w_{jk}/(1 - h_{jk}). \tag{4.27}$$

A family of general relative risk models that is intermediate in generality between (4.24) and (4.25) is given by

$$\lambda_{jk} = \exp(\alpha_j)r(\mathbf{x}_{jk}; \boldsymbol{\beta}),$$

where the relative risk function is specified by the power relation

$$\log r(\mathbf{x}_{jk}; \boldsymbol{\beta}) = \begin{cases} \dfrac{(1 + \mathbf{x}_{jk}\boldsymbol{\beta})^\rho - 1}{\rho} & , \rho \neq 0 \\ \log(1 + \mathbf{x}_{jk}\boldsymbol{\beta}) & , \rho = 0. \end{cases} \tag{4.28}$$

This yields the additive relative risk model (4.24) at $\rho = 0$ and the standard multiplicative model (4.10) at $\rho = 1$. Thomas, D.C. (1981) proposed another family $\mathrm{RR}(\mathbf{x}) = \exp(\rho\mathbf{x}\boldsymbol{\beta})\{1 + \mathbf{x}\boldsymbol{\beta}\}^{1-\rho}$, which also contains both additive and multiplicative forms. These two families, which specify how the exposure effects combine to yield a factor $r(\mathbf{x}; \boldsymbol{\beta})$, which then multiplies background, should be contrasted with the family (4.16),

Table 4.20 GLIM macros for fitting the relative risk model specified by equation (4.28) to grouped cohort data

```
$sub porr ! macros for power transform relative risks
!          MACRO PORR REQUIRES THE FOLLOWING INPUT DATA
!           r = NUMBER OF CASES (BINOMIAL NUMERATOR)
!           n = NUMBER OF CASES + CONTROLS (DENOMINATOR)
!          %l = POWER TRANSFORM USED IN RELATIVE RISK (%l = 0 FOR LOG)
!         strt = FACTOR WITH %i LEVELS THAT CONTAINS STRATUM INDICATORS
!          %i = NUMBER OF LEVELS FROM STRATUM INDICATOR STRT
!           b = INITIAL VALUES FOR PARAMETERS (LENGTH %l + 8)
!   x1 ... x8 = REGRESSION VARIABLES CODED TO HAVE POSITIVE RELATIVE RISK
                CODE Xl = 0 FOR THOSE THAT ARE NOT TO BE FITTED.
!
!          ON EXIT THE FOLLOWING QUANTITIES ARE AVAILABLE
!
!           p = PREDICTED PROBABILITY OF 'BEING A CASE'
!           h = DIAGONAL TERMS FROM 'HAT' MATRIX
!          cs = STANDARDIZED RESIDUALS (CHI-SQUARE TYPE)
!        %vc = MATRIX OF VARIANCES AND COVARIANCES OF ESTIMATES
!
$mac ftnl ! macro to fit nonlinear relative risk model
$cal %k = 10 : %c = 0.0001 ! set convergence criteria
$err n !
$wei w $yvar y $while %k porr $dis e $ext %vl $cal h = %vl*w !
$cal cs = (r - n*p)/%sqrt(w) : %t = %cu(cs*cs) !
$cal %u = 2*%cu(r*%log(r/(n*p)) + (n - r)*%log((n - r)/(n*(1 - p)))) !
$pri 'chi-square' %t 'deviance' %u $ !
$del %pe y w %fv z1 z2 z3 z4 z5 z6 z7 z8 xb th db %vl $$endmac !
$mac porr ! rr(x) = ((1 + xb)**%l - 1)/%l
$cal xb = b(%i + 1)*x1 + b(%i + 2)*x2 + b(%i + 3)*x3 + b(%i + 4)*x4 + b(%i + 5)*x5
 + b(%i + 6)*x6 + b(%i + 7)*x7 + b(%i + 8)*x8 !
$cal xb = %if(%le(xb, 0), 0.0001, xb) !
$cal %a = 1 + %eq(%l, 0) $switch %a pow log $
$cal xb = (1 + xb)**(%l - 1) !
$cal p = %exp(th) : p = p/(1 + p) !
$cal w = n*p*(1 - p) : y = (r - n*p)/w !
$cal z1 = x1*xb : z2 = x2*xb : z3 = x3*xb : z4 = x4*xb : z5 = x5*xb : z6 = x6*xb !
$cal z7 = x7*xb : z8 = x8*xb !
$sca 1 !
$fit strt - %gm + z1 + z2 + z3 + z4 + z5 + z6 + z7 + z8 $ext %pe $cal db = %pe : b = b + db !
$pri %k 'estimates' b !
$use cchk ! check for convergence
$$endmac
$mac pow $cal th = b(strt) + ((1 + xb)**%l - 1)/%l $$endmac !
$mac log $cal th = b(strt) + %log(1 + xb) $$endmac !
$mac cchk ! convergence check
$cal db = %if(%le(db, 0), -db, db)/b !
$cal db = %if(%le(db, %c), 0, 1) : %t = %cu(db) !
$cal %k = %k - 1 : %k = %if(%le(%t, 0), 0, %k) $$endmac !
$return
```

which specifies how exposure and background effects combine. Table 4.20 contains a series of GLIM macros, based on the general iterated least-squares methodology just described, that were used to fit the class of models (4.28) in the illustrative examples. The multiplicative model with $\rho = 1$ is easily fitted using standard features of GLIM, and convergence is guaranteed. A recommended procedure is to fit this model to get starting values of $\boldsymbol{\beta}^* = (\alpha, \boldsymbol{\beta})$ for use with nearby values of ρ, say $\rho = 0.9$. One then uses the $\boldsymbol{\beta}^*$ values obtained at convergence with $\rho = 0.9$ to start the procedure with $\rho = 0.8$, and so on until the additive relative risk model $\rho = 0$. However, due to the general problems with additive and other nonmultiplicative relative risk models mentioned above, it may prove impossible to implement this procedure with some data sets once ρ decreases beyond a certain point. Comparison of deviances for various values of ρ allows one to judge which (if either) of the additive or multiplicative relative risk models provides a reasonable description of the data, just as in Example 4.2. Thompson and Baker (1981) describe an alternative methodology for fitting models with composite link functions, which may be implemented using the OWN feature of GLIM to fit (4.28). Pierce *et al.* (1985) have developed a flexible program to fit models in which the rates are expressed as a sum of products of multiplicative and additive terms.

Example 4.6

In order to illustrate the fitting of the additive relative risk model, we consider another set of data from the British doctors study (Doll & Peto, 1978). Table 4.21 presents numbers of lung cancer deaths and

Table 4.21 Numbers of lung cancers (O) and person–years of observation (P–Y) by age and smoking level among British male doctors[a]

No. of cigarettes smoked per day	Average number smoked		Age in years							
			40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79
0	0	O	0	0	1	2	0	0	1	2
		P–Y	17846.5	15832.5	12226	8905.5	6248	4351	2723.5	1772
1–4	2.7	O	0	0	0	1	1	0	1	0
		P–Y	1216.0	1000.5	853.5	625	509.5	392.5	242.0	208.5
5–9	6.6	O	0	0	0	0	1	1	2	0
		P–Y	2041.5	1745	1562.5	1355	1068	843.5	696.5	517.5
10–14	11.3	O	1	1	2	1	1	2	4	4
		P–Y	3795.5	3205	2727	2288	1714	1214	862	547
15–19	16.0	O	0	1	4	0	2	2	4	5
		P–Y	4824	3995	3278.5	2466.5	1829.5	1237	683.5	370.5
20–24	20.4	O	1	1	6	8	13	12	10	7
		P–Y	7046	6460.5	5583	4357.5	2863.5	1930	1055	512
25–29	25.4	O	0	2	3	5	4	5	7	4
		P–Y	2523	2565.5	2620	2108.5	1508.5	974.5	527	209.5
30–34	30.2	O	1	2	3	6	11	9	2	2
		P–Y	1715.5	2123	2226.5	1923	1362	763.5	317.5	130
35–40	38.0	O	0	0	3	4	7	9	5	2
		P–Y	892.5	1150	1281	1063	826	515	233	88.5

[a] From Doll and Peto (1978)

Table 4.22 Results of fitting several relative risk models to the data on lung cancer in British doctors: internal estimation of age effects[a]

Model no.	Equation for relative risk (RR) or excess risk (ER) as a function of daily no. of cigarettes (x)	Degrees of freedom	Deviance	Parameter estimate	Standard error
1	Separate RR each dose group	56	45.74	(see Fig. 4.6)	
2	$RR = \exp(\beta x)$	63	68.91	0.0853	0.0063
3	$RR = \exp(\beta x + \gamma x^2)$	62	51.87	0.1801	0.0263 (β)
				−0.00226	0.00059 (γ)
4	$RR = (1 + x)^{\beta}$	63	55.87	1.187	0.123
5	$RR = 1 + \beta x$	63	58.36	1.130	0.510
6	$RR = 1 + \beta x + \gamma x^2$	62	51.03	0.4105	0.2880 (β)
				0.0237	0.0116 (γ)
7	Separate ER each dose group	56	184.7		
8	$ER = \beta x$	63	205.8	4.4×10^{-5}	0.6×10^{-5}

[a] From Breslow (1985b)

approximate person-years denominators classified by age and number of cigarettes smoked per day. Data for ages 80 and above were excluded from consideration, since the diagnosis is often uncertain at such advanced ages, while data for persons who reported smoking more than 40 cigarettes per day were excluded on the grounds of being unreliable and uncharacteristic. This latter exclusion, made also by Doll and Peto (1978), has a substantial impact on the dose-response analyses and has been the subject of some controversy.

Table 4.22 presents the results of fitting a variety of models to these data. In addition to the smoking parameters, estimates of which are shown in the table, each requires estimation of eight α_j parameters to represent the effects of age. The first four models are multiplicative. Smoking is treated qualitatively in model 1, with a separate relative risk being estimated for each smoking level. In models 2 and 3, the quantitative dose variable x = 'average number of cigarettes smoked per day' and then its square are introduced into the exponential term. Model 4 is the power relative risk model specified by equation (4.23), and models 5 and 6 are additive relative risk models as specified by (4.24).

Except for model 2, all the relative risk models (models 1–6) fit the observed data reasonably well. Model 4 fits best among those that require only a single parameter to describe the relative risk. The fact that a quadratic term significantly improves the fit of the additive relative risk model ($\chi^2_1 = 58.36 - 51.03 = 5.35$; $p = 0.02$) was interpreted by Doll and Peto (1978) as consistent with the notion that both an early and a late stage in the carcinogenic process are affected by cigarette smoke (see Chapter 6).

Figure 4.6 shows the relative risks estimated from four of the models. By definition, all relative risks are constrained to equal unity at zero dose. However, since most lung cancer deaths occur among smokers, the regression coefficients are largely determined by a comparison of rates for different classes of smokers, rather than by a comparison of smokers with nonsmokers. The fact that nonsmokers form the baseline category thus explains the apparently aberrant behaviour of the estimated relative risk curve for the power model. Were a more typical category used as a baseline, say, smoking of 20 cigarettes per day, all the curves would pass through unity at that point and would appear to be in better harmony. See the parallel discussion in §6.9 of Volume 1.

Certain drawbacks of the additive relative risk model are evident from Table 4.22. The standard errors for the regression coefficients are quite large in comparison with those for the multiplicative model, to the extent that t statistics of the form $t = \hat{\beta}/\mathrm{SE}(\hat{\beta})$ seriously understate the true statistical significance of the smoking effect. The t statistics for the multiplicative models 2 and 4 are $t = 0.0853/0.0063 = 13.5$ and $t = 1.187/0.123 = 9.7$, each highly significant, while that for model 5 is only $t = 1.130/0.510 = 2.2$. The contrast

Fig. 4.6 Three relative risk (RR) functions fitted to data on lung cancer rates from the British doctors study

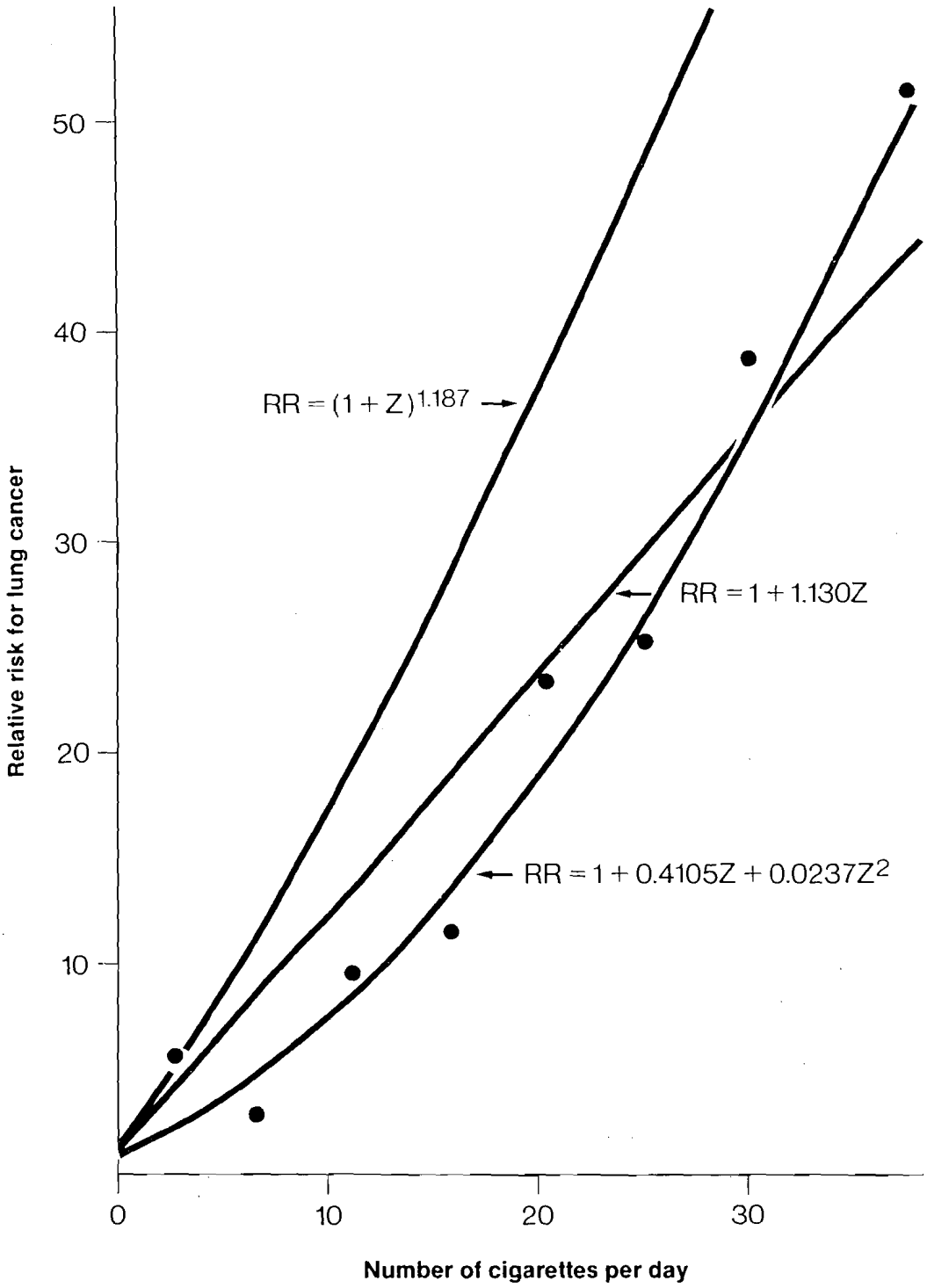
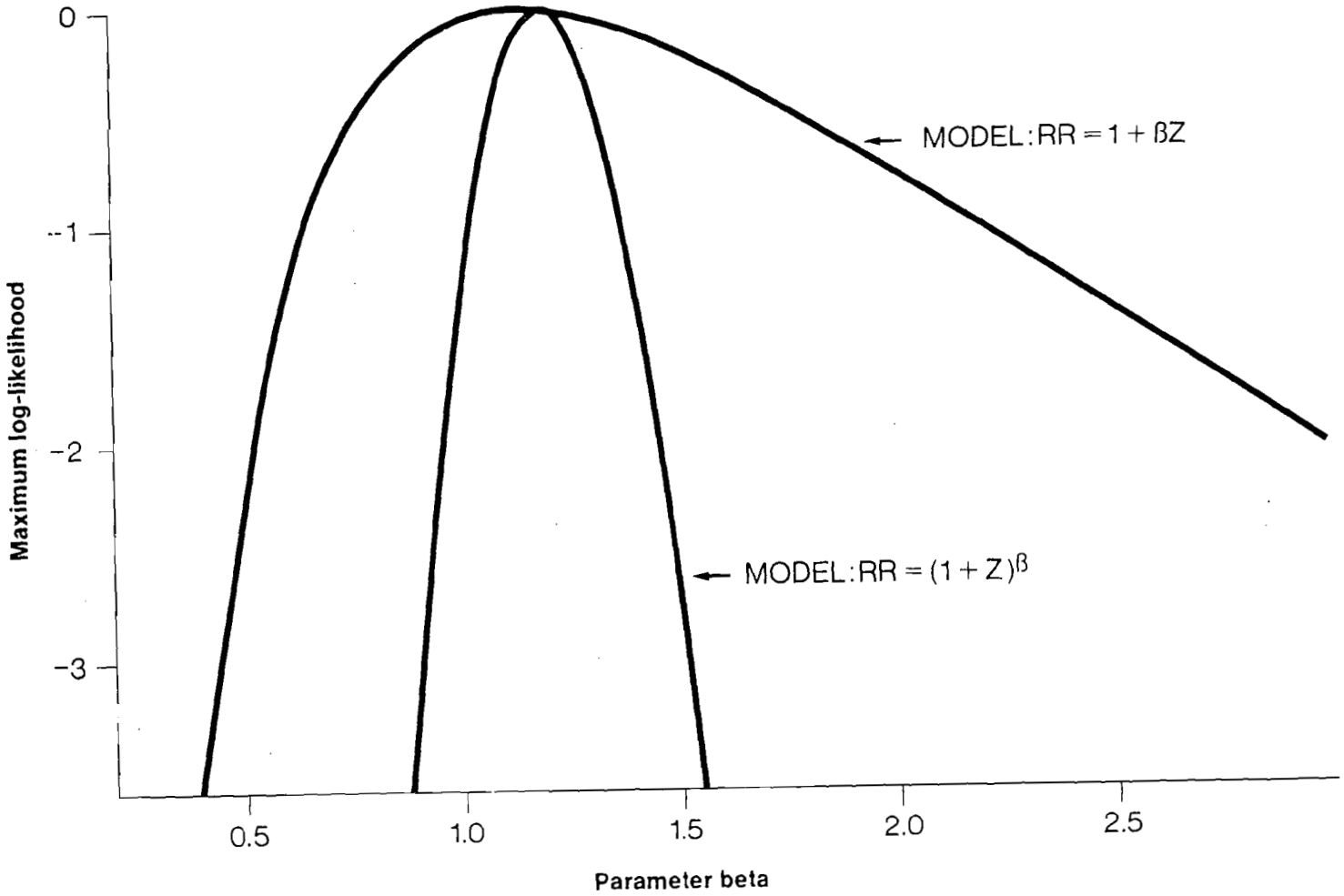


Fig. 4.7 Maximum (profile) log-likelihood functions for the additive and multiplicative relative risk (RR) models fitted to lung cancer rates from the British doctors study



between the maximum log-likelihood functions for the multiplicative and additive relative risk models (Fig. 4.7) is also striking. The β parameter in the additive relative risk model is less well determined, as indicated by the flatter log-likelihood, and there is substantial skewness. These problems may be largely overcome, however, by reparametrizing the model as $\lambda_{jk} = \alpha_j(1 + e^{\beta x_k})$ (Thomas, D.C., 1981), or as discussed by Barlow (1986).

The last two models in Table 4.21 are additive models. Model 7 has the form $\lambda_{jk} = \alpha_j + \beta_k$, where a separate excess risk β_k is estimated for each dose category. Model 8 is $\lambda_{jk} = \alpha_j + \beta x_k$, where the excess risk is assumed to be linear in dose. Neither fits the observed data at all well. In fact, as soon as one moves much away from the relative risk models 1–6, for example, by considering the power family (4.16) with ρ departing slightly from 0, the goodness-of-fit declines substantially. Thus, although there may be some doubt about the specific form of relative risk as a function of dose, smoking does appear to act multiplicatively on the age-specific rates.

(*d*) *Incorporating external standard rates*

External standard rates λ_j^* are incorporated into the additive relative risk model by writing it in the form

$$\begin{aligned} \lambda_{jk} &= \theta\lambda_j^*\{1 + \mathbf{x}_{jk}\boldsymbol{\beta}\} \\ &= \theta\lambda_j^* + \lambda_j^*\mathbf{x}_{jk}\boldsymbol{\psi}, \end{aligned} \tag{4.29}$$

where $\boldsymbol{\beta} = \boldsymbol{\psi}/\theta$ is the parameter of interest. This is formally equivalent to the additive model (equation (4.16) with $\rho = 1$), except that there is only one stratum parameter θ and all the regression variables are pre-multiplied by the known rates λ_j^*. Thus, it may be fitted directly in GLIM without recourse to the specialized macros given in Table 4.20.

Although there are fewer parameters to estimate, (4.29) has the same drawbacks as (4.24) with regard to instability of the β coefficients. Indeed, it is clear from the relation $\beta = \psi/\theta$ that much of the instability in this model is due to the extremely high dependence between the estimated relative risks and the estimates of the baseline rates, or between the relative risks and the scale factor θ used to adjust those rates.

Example 4.7

The same series of models considered in Example 4.6 was fitted to the British doctors data shown in Table 4.21, except that the baseline age-specific rates were assumed to be proportional to $\lambda_j^* = (t_j - 22.5)^{4.5} \times 10^{-11}$, where t_j is the midpoint of the jth age interval. Here, $t_j - 22.5$ represents the approximate duration of exposure to the putative carcinogen in the jth age group and the exponent 4.5 represents a compromise between five and six stages in the multistage theory of carcinogenesis (Doll & Peto, 1978). Thus, the external 'standard' rates are based on theoretical concepts, rather than on national vital statistics as in some earlier examples.

The results, shown in Table 4.23, are little different from those in Table 4.22, where the age effects were estimated directly from the data. One would expect that the standard errors of the regression coefficients of the smoking variables might be reduced somewhat, reflecting an increase in precision stemming from the stronger assumptions made about the background rates. Theoretical calculations (Stewart & Pierce, 1982; Breslow, 1985b) indicate that such an increase would be expected if age were a strong confounder, in the sense that average smoking levels changed markedly from one age group to the other. While there is some evidence for such confounding in these data, it is evidently not strong enough that knowledge of the background rates, at least up to a constant of proportionality, would contribute a significant advantage in terms of increased precision.

If the additive relative risk model is expressed in terms of the parameters θ and ψ, as in (4.29), we estimate $\hat{\theta} = 0.837 \pm 0.356$, $\hat{\psi} = 0.954 \pm 0.741$ and $\mathrm{Cov}(\hat{\theta}, \hat{\psi}) = -0.00686$. A test of the smoking effect is

Table 4.23 Results of fitting several relative risk models to the data on lung cancer in British doctors; age effects assumed proportional to $(\text{age}-22.5)^{4.5} \times 10^{-11}$

Model no.	Equation for relative risk (RR) or excess risk (ER) as a function of daily no. of cigarettes (x)	Degrees of freedom	Deviance	Parameter estimate	Standard error
1	Separate RR each dose group	63	47.13		
2	$RR = \exp(\beta x)$	70	70.29	0.0854	0.0063
3	$RR = \exp(\beta x + \gamma x^2)$	69	53.26	0.1802	0.0261 (β)
				−0.00226	0.00059 (γ)
4	$RR = (1 + x)^{\beta}$	70	57.35	1.192	0.122
5	$RR = 1 + \beta x$	70	60.00	1.141	0.516
6	$RR = 1 + \beta x + \gamma x^2$	69	52.43	0.409	0.286 (β)
				0.0239	0.0116 (γ)
7	Separate RR each dose group	63	200.0		
8	$ER = \beta x$	70	225.2	4.2×10^{-5}	0.6×10^{-5}

From Breslow (1985b)

thus given by $t = \hat{\psi}/\text{SE}(\hat{\psi}) = 12.9$, of the same order of magnitude as with the multiplicative models. However, the test based on $\hat{\beta} = \hat{\psi}/\hat{\theta} = 1.141$ divided by its standard error $\{\text{Var}(\hat{\theta})\hat{\beta}^2 - 2\,\text{Cov}(\hat{\theta}, \hat{\psi})\hat{\beta} + \text{Var}(\hat{\psi})\}^{1/2}/\hat{\theta} = 0.516$ yields a t statistic of only 2.2. This suggests that a large part of the instability in $\hat{\beta}$ in the additive relative risk model is its high correlation with parameter estimates (here $\hat{\theta}$) that represent the background rates. Further confirmation of this interpretation is given in Figure 4.8, which shows contour plots of the deviances obtained by varying the two parameters in the model equation (4.29). Although the minimum deviance of 60.0 occurs at $\hat{\beta} = 1.141$ (Table 4.23, model 5), nearly identical fits are obtained for a wide combination of parameter values (θ, β). The corresponding Figure 4.9 for the multiplicative model (Table 4.12, model 4) shows that, while there is still a strong dependence between the parameters representing relative risk and background, it is not so extreme as to lead to serious instability.

(e) *General risk functions for proportional mortality*

The relative risk models considered for proportional mortality analyses in §4.7 may also be generalized using the techniques of this section. In place of (4.19) we suppose

$$\begin{aligned} \lambda_{jk} &= \exp(\alpha_j) r(\mathbf{x}_{jk}; \boldsymbol{\beta}) \\ \nu_{jk} &= \exp(\gamma_j), \end{aligned} \tag{4.30}$$

where $r(x; \beta)$ denotes the general relative risk function for the cause of interest and where we have explicitly assumed that death rates for other causes are not affected by the exposures. The probabilities p_{jk} of 'being a case' then satisfy

$$\text{logit}\, p_{jk} = (\alpha_j - \gamma_j) + \log\, r(\mathbf{x}_{jk}; \boldsymbol{\beta}). \tag{4.31}$$

As shown above, a flexible and convenient family of models for the log relative risks (although by no means the only one that could be suggested for this purpose) is the

Fig. 4.8 Contour plot of deviances (G^2) when fitting the additive relative risk model with external standard rates to lung cancer rates from the British doctors study

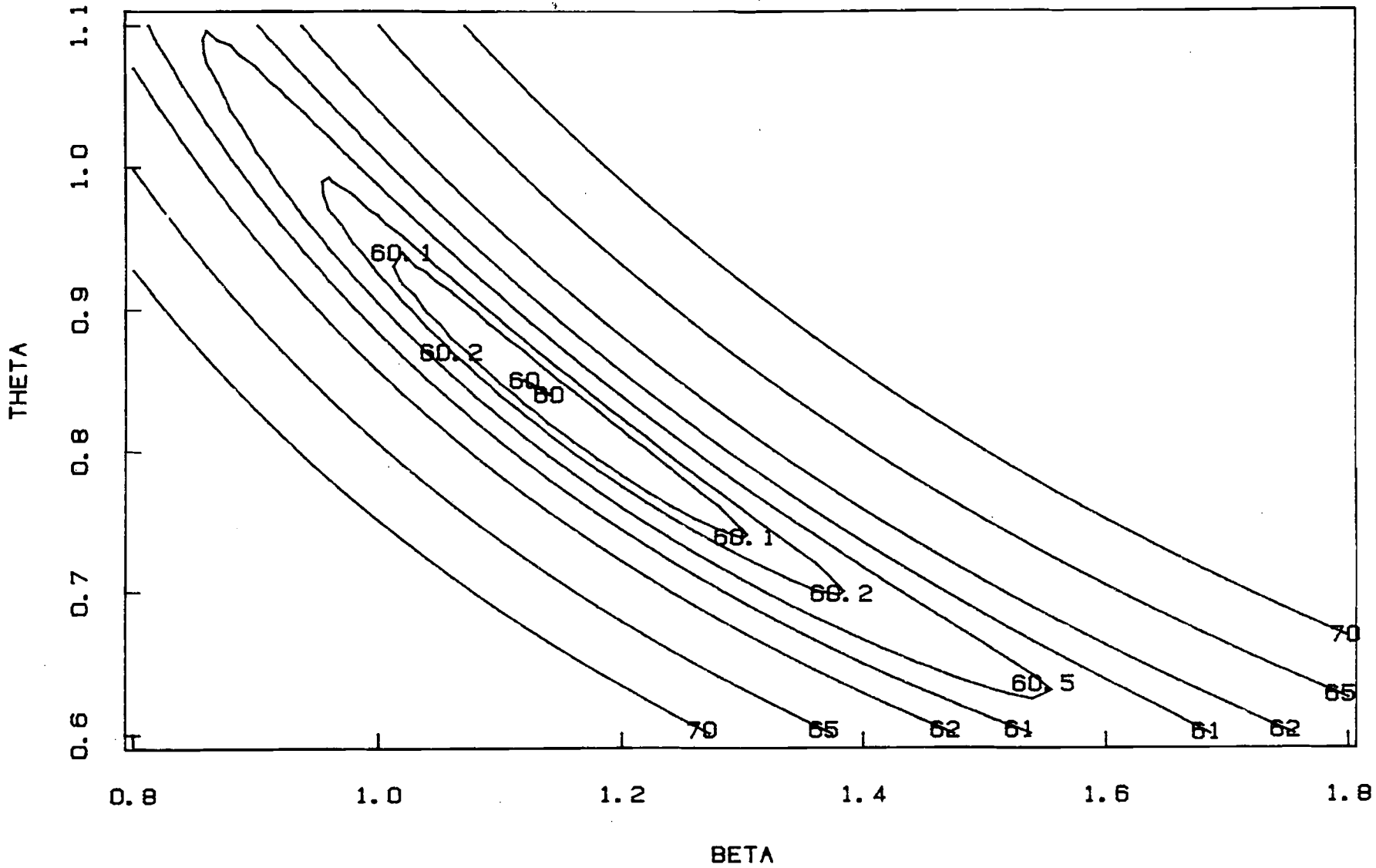
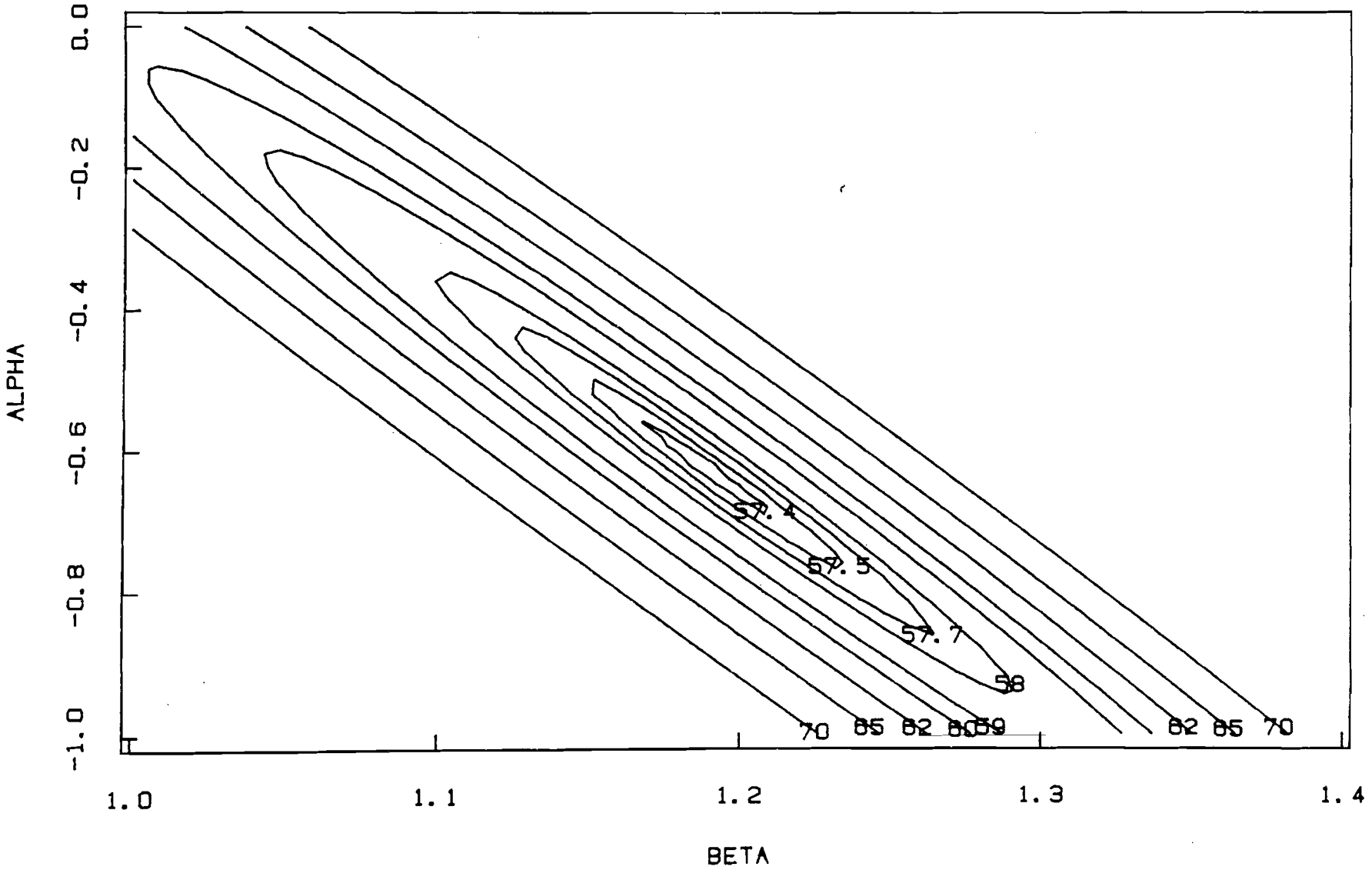


Fig. 4.9 Contour plot of deviances (G^2) when fitting the multiplicative model with external standard rates and log transform of smoking to lung cancer rates from the British doctors study



family

$$\log r(\mathbf{x};\boldsymbol{\beta}) = \frac{(1+\mathbf{x}\boldsymbol{\beta})^{\rho}-1}{\rho},$$

which contains both additive ($\rho \to 0$) and multiplicative ($\rho = 1$) relative risk functions as special cases. Breslow and Storer (1985) illustrate the fitting of such general relative risk functions to grouped data from actual case-control studies. The same techniques can be used in proportional mortality analyses.

4.10 Fitting relative and excess risk models to grouped data on lung cancer deaths among Welsh nickel refiners

Appendix ID presents a detailed discussion of the background and design of the study of Welsh nickel refinery workers. Summary data on nasal sinus cancer deaths (Appendix VI) were considered briefly in Example 4.1 in order to illustrate some features of the fitting of multiplicative models to grouped data. Published data from this study provided us in Chapter 3 with examples of the use of internal standardization. With the approval of Kaldor *et al.* (1986), we undertake in this section a more comprehensive analysis of grouped data on lung cancer deaths in order to contrast the results obtained with relative and excess risk models. In the next chapter, continuous variable modelling techniques are applied to the study of rates of nasal sinus cancer deaths that occurred among these same workers.

(*a*) *Basic data and summary statistics*

Table 4.24 was compiled by Peto, J. *et al.* (1984) to summarize the mortality experience through 1981. The excess mortality was due largely to nasal sinus and lung cancers and was essentially confined to the 679 men employed before 1925, to whom attention is henceforth confined. Appendix VIII lists basic data for each of the 679 men that were used for all the grouped and continuous variable analyses reported in this

Table 4.24 Mortality experiences (O, observed; E, expected) of Welsh nickel refiners[a]

Period first employed	Number of men		Cancers			Other causes			All causes
			Lung	Nasal sinus	Other	Circulatory disease	Respiratory disease	Other	
Before 1925	679	O	137	56	67	220[b]	63	60	603
		E	26.86	0.21	59.44	194.76	62.39	75.74	419.38
1925–1929	97	O	11	0	11	26[b]	13	14	75
		E	5.48	0.03	8.08	26.19	8.46	9.04	57.28
1930–1944	192	O	11	0	16	58	13	12	110
		E	9.13	0.05	12.90	42.43	12.92	11.70	89.12

[a] From Peto, J. *et al.* (1984)
[b] Including one death in which nasal sinus cancer was an underlying cause

monograph. There are slight differences between this data set and that analysed by Peto, J. *et al.* (1984), Kaldor *et al.* (1986) and others, due to the continual process of data editing. Thus, the person-years of observation, expected numbers of deaths and relative risks shown in Tables 4.24 and 6.8 and in our own summaries (e.g., Table 4.25) differ very slightly. However, these differences have no material effect on the results or interpretation.

Six basic pieces of information are available for each subject: (i) ICD (7th revision) cause of death; (ii) exposure, defined as the number of years worked in one of seven 'high-risk' job categories prior to the start of follow-up (see below); (iii) date of birth; (iv) age at initial employment; (v) age at start of follow-up; and (vi) age at death for those who died, age last seen for those lost to observation, or age at end of study for those withdrawn alive. Nasal sinus cancer deaths are coded 160 under the 7th ICD revision, and lung cancer deaths are 162 or 163. Further dates of interest, such as date entered follöw-up and date of initial employment, are obtained by adding the corresponding ages to the date of birth.

The nasal sinus cancer, lung cancer and total (all causes) death rates for England and Wales by five-year intervals of age and calendar time, listed in Appendix IX, were used to compute the expected numbers of deaths and the values of an age-dependent covariable, consisting of the standard death rate for each subject, at specified points in time.

(*b*) *Construction of the exposure index*

Company records were used to classify each year of an individual's employment into one of ten categories, depending on the area of the plant in which he worked on 1 April of that year. Such data were available for 82% of the 9354 calendar years during which the 679 subjects were employed prior to 1925. Kaldor *et al.* (1986) used a synthetic case-control approach to analyse the relation between work area and respiratory (nasal sinus and lung) cancer risk. They identified five exposure categories that appeared to be significantly related to the risk of both cancers: calcining I, calcining II, copper sulphate, nickel sulphate and furnaces. (See Table 6.7.) On this basis, they developed an exposure index equal to the number of calendar years employed in these categories. A contribution of a half rather than a full year was given for the first and last calendar year of such employment. In contrast to the Montana study, there was no overlap of exposure and follow-up periods and hence no change in the exposure index with follow-up. This simplified the analysis considerably.

Due to the circularity involved in construction of the exposure index, the excess risks may be overstated slightly. Another possible deficiency is that the index does not account for the time or age at which 'high-risk' exposures were received. Any difference between high-risk exposures received during 1905–1909 and those received between 1920 and 1924 is ignored. Furthermore, the use of date of initial employment to represent the start of exposure may obscure the fact that relevant exposures were primarily received in high-risk areas. One might consider an analysis of two exposure duration variables – years since initial employment *and* years since initial employment in a high-risk area. However, due to the undoubtedly high correlation between them,

estimation of their separate effects would be problematic. Hence, time since first employment is used as the only time-dependent variable in the ensuing analyses.

(*c*) *Grouping of data for analysis*

From the data in Appendix VIII we constructed a four-dimensional table of observed numbers of lung cancer deaths, person-years of observation, and expected numbers of lung cancer deaths, and likewise for nasal sinus cancer. The dimensions were (i) age at first employment (AFE) in four levels; (ii) calendar year of first employment (YFE) in four levels; (iii) the exposure index (EXP) in five levels; and (iv) time since first employment (TFE) in five levels. The calculation used Clayton's algorithm (Appendix IV) to combine the 679 original records with the national death rates for England and Wales. At one point it was necessary to add two more dimensions to the table, namely, current age and calendar year in the quinquinquennia for which the national rates were available. Person-years in each cell were multiplied by the corresponding standard rate and then summed to give expected numbers of lung cancer deaths. Only 242 of the $4 \times 4 \times 5 \times 5 = 400$ cells in the four-dimensional table had some person-years of observation time available. The data for these 242 cells are presented in Appendix VII so that the reader can more easily verify our results.

(*d*) *Fitting the relative risk model*

Table 4.25 shows the person-years and observed and expected numbers accumulated for each factor level. The sixth column of the table presents estimated relative risks (ratios of SMRs) for each factor, adjusted for the remaining three factors. These were estimated from a multiplicative model (equation (4.18)) that incorporated the standard rates and 14 binary regression variables to represent the simultaneous effects of the four factors. An overall SMR of 8.92 was estimated for the baseline category, namely for the period up to 20 years since date of hire for workers hired under 20 years of age before 1910 with no time spent in a high-risk job. The lung cancer relative risk increased fourfold with increasing exposure, but declined markedly as TFE advanced beyond 20 years. The smaller changes in the SMR with age and year of initial employment were not statistically significant (Table 4.26).

(*e*) *Fitting the excess risk model*

Due to the ageing of the cohort and the secular increase in cigarette smoking, the national rates used to determine the SMRs were themselves climbing rapidly with increasing follow-up. Thus, it is unclear from the decline in the SMRs with TFE what the temporal evolution of absolute excess risk may be. In order to investigate this question, Kaldor *et al.* (1986) employed a model for excess risk that had been proposed earlier by Brown and Chu (1983), namely,

$$\lambda_{jk} = \lambda_j^* + \exp(\alpha + \mathbf{x}_k\boldsymbol{\beta}). \tag{4.32}$$

Here, $\exp(\alpha)$ represents the excess mortality rate for someone with a standard set of

Table 4.25 Fitting of relative and excess risk models to data on lung cancer mortality among Welsh nickel refiners

Risk factor	Level	Person–years at risk	No. of lung cancer deaths		Relative risk (ratio of SMRs)	Excess mortality ratio (EMR)
			Observed	Expected[a]		
Age at	<20 years	3089.2	13	5.58	1.0	1.0
first	20–27.5	7064.9	72	13.30	1.43	2.78
employment	27.5–35.0	3711.9	41	6.25	1.23	2.86
(AFE)	35.0+	1364.9	11	1.75	0.91	2.64
Year of	1900–1909	1951.0	23	2.84	1.0	1.0
first	1910–1914	2904.5	39	4.52	1.08	1.37
employment	1915–1919	2294.0	13	4.12	0.62	0.99
(YFE)	1920–1924	8081.3	62	15.39	0.73	1.70
Exposure	0–	7738.8	42	14.91	1.0	1.0
index	0.5–4.0	4905.1	50	8.32	1.83	2.41
(EXP)	4.5–8.0	1716.9	27	2.47	2.95	4.19
	8.5–12.0	601.2	12	0.85	3.54	5.04
	12.5+	269.9	6	0.34	4.03	5.87
Time since	0–19 years	2586.1	6	0.56	1.0	1.0
first	20–29	4777.5	35	3.15	0.76	2.87
employment	30–39	4329.4	55	7.60	0.48	5.02
(TFE)	40–49	2461.4	31	9.20	0.22	4.77
	50+	1076.4	10	6.37	0.11	2.11
Baseline SMR:					8.92	
Baseline excess mortality (per 100 000 person–years)						30.0
Chi-square goodness-of-fit (deviance; 227 degrees of freedom)					195.4	194.8

[a] Based on rates for England and Wales by age and calendar year (Appendix IX)

covariable values ($\mathbf{x}_k = \mathbf{0}$), and $\exp(\mathbf{x}_k\boldsymbol{\beta})$ represents the excess mortality ratio (EMR), i.e., the factor by which the specific exposures modify the excess rate.

The model defined by (4.32) may be fitted easily with the GLIM OWN facility for user-defined models, just as (4.18) is fitted using standard features of the program. Table 4.27 lists the GLIM commands needed to read the 242 data records from

Table 4.26 Evaluating the significance of variations in the SMR and EMR for each risk factor: lung cancer mortality among Welsh nickel refiners

Risk factor[a]	Degrees of freedom	Effect on relative mortality (SMR difference)		Effect on excess mortality (EMR difference)	
		Chi-square	*p* value	Chi-square	*p* value
AFE	3	2.96	0.40	7.35	0.06
YFE	3	3.77	0.29	3.34	0.34
EXP	4	21.25	0.0003	24.26	0.0001
TFE	4	42.42	<0.0001	17.25	0.002

[a] AFE, age at first employment; YFE, year of first employment; EXP, exposure index; TFE, time since first employment

Table 4.27 GLIM commands needed to produce the results shown in Table 4.25[a]

```
$UNITS 242 $DATA AFE YFE EXP TFE CASES PYR EXPT $
$FORMAT
(4(X, F2.0), 11X, F3.0, 8X, 2(X, F12.6))
$C READ IN DATA FROM FORTRAN UNIT 1 – APPENDIX VII
$DINPUT 1 80$
$C DECLARE OFFSET TO BE LOGARITHM OF EXPECTED NUMBERS OF CASES
$CAL EXPT = EXPT/1000 $
$CAL LEX = %LOG(EXPT) $OFF LEX $
$YVAR CASES $ERR P $
$FAC 242 AFE 4 YFE 4 EXP 5 TFE 5 $
$C FIT MULTIPLICATIVE MODEL
$FIT AFE + YFE + EXP + TFE $
$DIS M E $
$C EXTRACT PARAMETER ESTIMATES AND CONVERT INTO RELATIVE RISKS
$EXT %PE $CAL RR = %EXP(%PE) $LOOK RR $
$C NOW CONTINUE WITH EMR MODEL
$OFF $
$MAC M1 $CAL %FV = %EXP(%LP)*PYR + EXPT $$ENDMAC
$MAC M2 $CAL %DR = 1./(%FV – EXPT) $$ENDMAC
$MAC M3 $CAL %VA = %FV $$ENDMAC
$MAC M4 $CAL %YV = %IF(%LT(%YV, .5), .0000001, %YV) $
$CAL %DI = 2.*(%YV*%LOG(%YV/%FV) – (%YV – %FV)) $$ ENDMAC
$OWN M1 M2 M3 M4 $FIT . $REC 10 $FIT . $
$DIS M E $EXT %PE $CAL RR = %EXP(%PE) $LOOK RR $
$STOP
```

[a] Adapted from Kaldor *et al.* (1986)

Appendix VII and produce the SMR and EMR estimates shown in Table 4.25. The excess risk was estimated to be approximately 30 lung cancer cases per 100 000 person-years for workers in the baseline category. It also increased sharply with the exposure index. By contrast to the pattern in the relative risk, however, the excess risk increased to a maximum some 40 years from date of hire and subsequently declined.

Brown and Chu (1983) note that one will sometimes wish to adjust the standard rates λ_j^* used in modelling the excess risk in order to account for the healthy worker selection bias or other systematic departures of baseline mortality rates in the study group from the national averages. They suggest $\lambda_{jk} = \theta\lambda_j^* + \exp(\alpha + \mathbf{x}_k\boldsymbol{\beta})$ as a generalization of (4.32) and arbitrarily set $\theta = 0.8$ or $\theta = 1.2$ in order to gauge the sensitivity of the $\boldsymbol{\beta}$ parameter estimates to variation in the assumed background rates. We used the AMFIT program of Pierce *et al.* (1985) to estimate θ by maximum likelihood and found $\hat{\theta} = 1.087$. Since there was scarcely any improvement in fit over the model with $\theta = 1$, it appears that the national rates do a reasonable job of representing background lung cancer mortality for the Welsh cohort.

Although the SMR and EMR models happen to fit this particular set of data equally well, they lead to markedly different estimates of lifetime risk for typical workers. Kaldor *et al.* (1986) estimated the lifetime (to age 85) probability of lung cancer for light smokers who were born in 1900, who started work at the nickel refinery in 1920,

and who accumulated an exposure index of 10. The probability was 0.27 under the multiplicative model and 0.58 under the additive one. For a heavy smoker, the estimated lifetime probability was 0.65 for the multiplicative model and 0.61 for the additive one.

Section 6.6 presents a further discussion of these results in terms of the multistage theory of carcinogenesis.