6. UNCONDITIONAL LOGISTIC REGRESSION FOR LARGE STRATA

CHAPTER 6

UNCONDITIONAL LOGISTIC REGRESSION FOR LARGE STRATA

The elementary techniques described above for stratified analysis of case-control studies, and in particular the Mantel-Haenszel combined relative risk estimate and test statistic, have served epidemiologists well for over two decades. Most of the calculations are simple enough for an investigator to carry out himself, although this often means devoting considerable time to routine chores. Some of the boredom may be alleviated through the use of modern programmable calculators, for which the methods are ideally suited. By working closely with his data, examining them in tabular form, calculating relative risks separately for each stratum, and so on, the researcher can spot trends or inconsistencies he might not otherwise have noticed. Errors in the data may be discovered in this way, and new hypotheses generated.

Nevertheless there are certain limitations inherent in the elementary techniques that must be recognized. If many potentially confounding factors must be controlled simultaneously, a stratified analysis will ultimately break down. Individual strata simply become so large in number and small in size that many of them contain only cases or only controls. This means that substantial amounts of data are effectively lost from the analysis. There are similar limits on the number of categories into which continuous risk factors can be broken down for calculation of separate estimates of relative risk. It is desirable to leave them as continuous variables for purposes of interpolation and extrapolation. The inconsistencies arising from the selection of different levels of a variable to serve as baseline have already been noted, and while often relatively minor, these can be irritating. Limitations are likewise imposed on the extent to which one can analyse the joint effects of several risk factors. Perhaps even more important are the deficiencies in the elementary methods for evaluating interactions among risk and nuisance variables. The usual tests are notoriously lacking in statistical power against patterns of interaction which one might well expect to observe in practice. Other than calculating a separate estimate for each stratum, no provision is made for incorporating such interactions into the estimates of relative risk.

Access to high-speed computing machinery and appropriate statistical software removes these limitations and opens up new possibilities for the statistical analysis of case-control data. By entering a few simple commands into a computer terminal, the investigator can carry out a range of exploratory analyses which could take days or weeks to perform by hand, even with a programmable calculator. He has a great deal of flexibility in choosing how variables are treated in the analysis, how they are categorized, or how they are transformed. The possibilities for multivariate analysis are virtually limitless. Such methods should, of course, be used in conjunction with tabular presentation of the basic data. Liberal use of charts and graphs to represent the results of the analyses is also recommended.

The basic tool which allows the scope of case-control study analysis to be thus broadened is the linear logistic regression model. Here we introduce the logistic model as a method for multivariate analysis of prospective or cohort studies, which reflects the historical fact that the model was specifically designed for, and first used with, such investigations. Its equal suitability for use in case-control investigations follows as a logical consequence. We replicate the stratified analyses of Chapter 4 using the modelling approach, and then extend these analyses by the inclusion of additional variables so as to illustrate the full power and potential of the method.

Unfortunately the level of statistical sophistication demanded from the reader for full appreciation of the modelling approach is more advanced than it has been in the past. While we have attempted to make the discussion as intelligible as possible for the non-specialist, familiarity with certain aspects of statistical theory, especially linear models and likelihood inference, will undoubtedly facilitate complete understanding.

6.1 Introduction to the logistic model

Whether using the follow-up or case-control approach to study design, cancer epidemiologists typically collect data on a number of variables which may influence disease risk. Each combination of different levels of these variables defines a category for which an estimate of the probability of disease development is to be made. For example, we way want to determine the risk of lung cancer for a man aged 55 years who has worked 30 years as a telephone linesman and smoked 20 cigarettes per day since his late teens.

If a large enough population were available for study, and if we had unlimited time and money, an obvious approach to this problem would be to collect sufficient numbers of subjects in each category in order to make a precise estimate of risk for each category separately. Of course in the case-control situation these risk estimates would not be absolute, but instead would be relative to that for a designated baseline category. With such a vast amount of data there would be no need to borrow information from neighbouring categories, i.e., those having identical levels for some of the risk variables and similar levels for the remainder, in order to get stable estimates of risk.

Epidemiological studies of cancer, however, rarely even come close to this ideal. Often the greatest limitation is simply the number of cases available for study within a reasonable time period. While this number may be perfectly adequate for assessing the relative risks associated with a few discrete levels of a single risk factor, it is usually insufficient to provide separate estimates for the large number of categories generated by combining even a few more or less continuous factors. Thus we are faced with the problem of having to make *smoothed* estimates which do utilize information from surrounding categories in order to estimate the risks in each one.

Such smoothing is carried out in terms of a *model,* which relates disease risk to the various combinations of factor levels which define each risk category *via* a *mathematical formula.* The model gives us a simplified, quantitative description of the main features of the relationship between the several risk factors and the probability of disease development. It enables us to *predict* the risk even for categories in which scant information is available. Important features for the model to have are that it provide meaningful results, describe the observed data well and, within these constraints, be as simple as possible. In view of the discussion in Chapter 2, therefore, the *parameters* of any proposed model should be readily interpretable in terms of relative risk. The model should also allow relative risks corresponding to two or more distinct factors to be represented as the product of individual relative risks, at least as a first approximation.

A model which satisfies these requirements, indeed which has in part been developed specifically to meet them, is the *linear logistic model.* It derives its name from the fact that the *logit* transform of the disease probability in each risk category is expressed as a linear function of *regression* variables whose values correspond to the levels of exposure to the risk factors. In symbols, if P denotes the disease risk, the logit transform y is defined by

$$y = \text{logit } P = \log\left(\frac{P}{1-P}\right), \tag{6.1}$$

or, conversely, expressing P in terms of y,

$$P = \frac{\exp(y)}{1+\exp(y)}\,. \tag{6.2}$$

Since P/(1–P) denotes the disease odds, another name for logit is *log odds.* Cox (1970) develops the theory of logistic regression in some detail.

The simplest example of logistic regression is provided by the ubiquitous 2 × 2 table considered in § 2.8 and § 4.2. Suppose that there is but a single factor and two risk categories, exposed and unexposed, and let P_1 and P_0 denote the associated disease probabilities. According to the discussion in § 2.8 the key parameter, which is both estimable from case-control studies and interpretable as a relative risk, is the odds ratio

$$\psi = \frac{P_1Q_0}{P_0Q_1}\,.$$

Its logarithm, i.e., the log relative risk, may be expressed

$$\beta = \log\psi = \text{logit } P_1 - \text{logit } P_0$$

as the *difference* between two logits. Let us define a single binary regression variable x by x = 1 for exposed and x = 0 for unexposed. If we write P(x) for the disease probability associated with an exposure x, and $r(x) = P(x)Q_0/P_0Q(x)$ for the relative risk (odds ratio relative to x = 0), we have

$$\log r(x) = \beta x$$

or

$$\text{logit } P(x) = \alpha + \beta x, \tag{6.3}$$

where $\alpha = \text{logit } P_0$. There is a perfect correspondence between the two parameters α and β in the model and the two disease risks such that

$$P_1 = \frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)}$$

and

$$P_0 = \frac{\exp(\alpha)}{1 + \exp(\alpha)} .$$

The formulation (6.3) focuses on the key parameter, β, and suggests how to extend the model for more complex problems.

A more interesting situation arises when there are two risk factors A and B, each at an exposed (+) and unexposed (−) level (§ 2.6). The combined levels of exposure yield four risk categories with associated disease probabilities P_{ij}:

$$\begin{array}{c|c|c|}
 & \multicolumn{2}{c}{\text{Factor B}} \\
\text{Factor A} & + & - \\
\hline
+ & P_{11} & P_{10} \\
\hline
- & P_{01} & P_{00} \\
\hline
\end{array} . \tag{6.4}$$

Taking P_{00} as the baseline disease risk, there are three relative risks to be estimated, corresponding to the three odds ratios

$$\psi_{10} = \frac{P_{10}Q_{00}}{P_{00}Q_{10}} \approx r_A$$

$$\psi_{01} = \frac{P_{01}Q_{00}}{P_{00}Q_{01}} \approx r_B$$

and

$$\psi_{11} = \frac{P_{11}Q_{00}}{P_{00}Q_{11}} \approx r_{AB}.$$

Here r_A, r_B and r_{AB} are relative risks for single and joint exposures, relative to no exposure, as defined in § 2.6.

We are particularly interested in testing the multiplicative hypothesis $r_{AB} = r_A r_B$ under which the relative risk for exposure to A is independent of the levels of B or, equivalently, the relative risk for B is independent of exposure to A. Expressed in terms of the odds ratios this becomes

$$\psi_{11} = \psi_{10}\psi_{01}. \tag{6.5}$$

If the hypothesis appears to fit the observed data, we should be able to summarize the risks for the three exposure categories relative to the baseline category in two numbers, *viz* the estimated relative risks for factors A and B individually. Otherwise a separate estimate for each of the three exposure categories will be required. We considered in § 4.4 some *ad hoc* tests for the multiplicative hypothesis and suggested that the Mantel-

Haenszel formula be used to estimate the individual relative risks if the hypothesis were accepted.

Estimates and tests of the multiplicative hypothesis are simply obtained in terms of a logistic regression model for the disease probabilities (6.4). Define the binary regression variable $x_1 = 1$ or 0 according to whether a person is exposed to Factor A or not, and similarly let x_2 indicate the levels of exposure to Factor B. Variables such as x_1 and x_2, which take on 0–1 values only, are sometimes called *dummy* or *indicator* variables since they serve to identify different levels of exposure rather than expressing it in quantitative terms. Note that the product x_1x_2 equals 1 only for the double exposure category. Let us define $P(x_1,x_2)$ as the disease probability, and $r(x_1,x_2)$ as the relative risk (odds ratio) relative to the unexposed category $x_1 = x_2 = 0$. Then we can *re-express* the relative risks, or equivalently the probabilities, using the model

$$\log r(x_1,x_2) = \beta_1x_1+\beta_2x_2+\gamma x_1x_2$$

i.e.,

$$\text{logit } P(x_1,x_2) = \alpha+\beta_1x_1+\beta_2x_2+\gamma x_1x_2. \tag{6.6}$$

Since there are four parameters α,β_1,β_2 and γ to describe the four probabilities P_{ij}, we say that the model is completely *saturated.* It imposes *no constraints* whatsoever on the relationships between the four probabilities or the corresponding odds ratios. Thus we may solve equation (6.6) explicitly for the four parameters, obtaining

$$\alpha = \text{logit } P_{00}$$

as the logit transform of the baseline disease probability,

$$\beta_1 = \log \psi_{10}$$

and

$$\beta_2 = \log \psi_{01}$$

as the log relative risks for individual exposures, and

$$\gamma = \log\left(\frac{\psi_{11}}{\psi_{10}\psi_{01}}\right) \tag{6.7}$$

$$= \text{logit } P_{11} - \text{logit } P_{10} - \text{logit } P_{01} + \text{logit } P_{00}$$

as the *interaction* parameter. It is clear from (6.7) that $\exp(\gamma)$ represents the multiplicative factor by which the relative risk for the double exposure category differs from the product of relative risks for the individual exposures. If $\gamma > 0$, a *positive interaction,* the risk accompanying the combined exposure is greater than predicted by the individual effects; if $\gamma < 0$, a *negative interaction,* the combined risk is less. Testing the multiplicative hypothesis (6.5) is equivalent to testing that the interaction parameter γ in the logistic model is equal to 0.

If the hypothesis $\gamma = 0$ is accepted by our test criterion, we would consider fitting to the data the reduced three parameter model

$$\text{logit } P(x_1,x_2) = \alpha + \beta_1x_1 + \beta_2x_2, \tag{6.8}$$

which re-expresses the multiplicative hypothesis in logit terms. This model does impose constraints on the four disease probabilities P_{ij}. For example, since

$$\beta_1 = \log \psi_{10} = \log \frac{\psi_{11}}{\psi_{01}}$$

now represents the log relative risk for A *whether or not* exposure to B occurs, it would be estimated by combining information from the 2×2 tables

	Factor B + Factor A		Factor B – Factor A	
	+	–	+	–
Cases	a_1	b_1	a_2	b_2
Controls	c_1	d_1	c_2	d_2
Odds ratio	ψ_{11}/ψ_{01}		ψ_{10}	

Likewise the estimate of

$$\beta_2 = \log \psi_{01} = \log \frac{\psi_{11}}{\psi_{10}}$$

would combine information from both the tables

	Factor A + Factor B		Factor A – Factor B	
	+	–	+	–
Cases	a_1	a_2	b_1	b_2
Controls	c_1	c_2	d_1	d_2 .
Odds ratio	ψ_{11}/ψ_{10}		ψ_{01}	

The difference between the interpretation of β_1 in (6.6) and the same parameter in (6.8) illustrates that *the meaning of the regression coefficients in a model depends on what other variables are included.* In the saturated model β_1 represents the log relative risk for A at level 0 of B only, whereas in (6.8) it represents the log relative risk for A at both levels of B. Testing the hypothesis $\beta_1 = 0$ in (6.8) is equivalent to testing the hypothesis that Factor A has no effect on risk, against the alternative hypothesis that there is an effect, but one which does not depend on B. It makes little sense to test $\beta_1 = 0$ in (6.6), or more generally to test for main effects being zero in the presence of interactions involving the same factors. Models which contain interaction terms without the corresponding main effects correspond to hypotheses of no *practical* interest (Nelder, 1977).

The regression approach is easily generalized to incorporate the effects of more than two risk factors, or risk factors at more than two levels. Suppose that Factor B occurred at three levels, say 0 = low, 1 = medium and 2 = high. There would then be six disease probabilities

		Factor B	
Factor A	High	Medium	Low
Exposed	P_{12}	P_{11}	P_{10}
Unexposed	P_{02}	P_{01}	P_{00}

and five odds ratios

$$\psi_{ij} = \frac{P_{ij}\,Q_{00}}{P_{00}\,Q_{ij}},$$

where all risks are expressed relative to P_{00} as baseline. In order to identify the three levels of Factor B, two indicator variables x_2 and x_3 are required in place of the single x_2 used earlier. These are coded as follows:

		Factor B	
	High	Medium	Low
x_2 =	0	1	0
x_3 =	1	0	0

.

More generally, for a factor with K levels, K–1 indicator variables will be needed to describe its effects. With x_1 defining exposure to A as before, the saturated model with six parameters is written

$$\text{logit } P_{ij} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_{12} x_1 x_2 + \gamma_{13} x_1 x_3,$$

where the values of the x's are determined from the factor levels i and j. Now the multiplicative hypothesis

$$\psi_{ij} = \psi_{i0}\psi_{0j}$$

corresponds to setting both interaction parameters γ_{12} and γ_{13} to zero, in which case the coefficients β_2 and β_3 represent the log relative risks for levels 1 and 2 of Factor B as compared with level 0.

If instead there are three factors A, B and C each at two levels, the disease probabilities may be denoted

	Factor C +		Factor C −	
	Factor B		Factor B	
Factor A	+	−	+	−
+	P_{111}	P_{101}	P_{110}	P_{100}
−	P_{011}	P_{001}	P_{010}	P_{000}

.

Here there are seven odds ratios $\psi_{ijk} = \dfrac{P_{ijk}\,Q_{000}}{P_{000}\,Q_{ijk}}$ to be estimated. The fully saturated model may be written

$$\text{logit } P_{ijk} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_{12} x_1 x_2 + \gamma_{13} x_1 x_3 + \gamma_{23} x_2 x_3 + \delta_{123} x_1 x_2 x_3, \qquad (6.9)$$

where x_1, x_2 and x_3 are indicator variables which identify exposures to factors A, B and C, respectively. The last parameter δ_{123} denotes the *second order interaction* involving all three variables. It has several equivalent representations in terms of the odds ratios or disease probabilities. One of these, for example, is

$$\begin{aligned}\delta_{123} &= \log \frac{\psi_{111}}{\psi_{101}\psi_{011}} - \log \frac{\psi_{110}}{\psi_{100}\psi_{010}} \\ &= \text{logit } P_{111} - \text{logit } P_{101} - \text{logit } P_{011} + \text{logit } P_{001} \\ &\quad - \{\text{logit } P_{110} - \text{logit } P_{100} - \text{logit } P_{010} + \text{logit } P_{000}\},\end{aligned}$$

viz the difference between the AB interaction at level 1 of Factor C and that same interaction at level 0 of Factor C. Other representations would be the difference between the AC interactions at the two levels of B, or the difference between the BC interactions at the two levels of A.

The advantage of expressing the disease probabilities in an equation such as (6.9) is that the higher order interactions generally turn out to be negligible. This permits the relative risks for all the cells in the complete cross-classification to be estimated using a smaller number of parameters which represent the main multiplicative effects of the important risk factors plus occasional lower order interactions. By reducing the number of independent parameters which must be estimated from the data, we achieve the smoothing which was noted earlier to be one of the primary goals of the analysis. If high-order interactions are found to be present, this alerts us to the fact that risk depends in a complicated way on the constellation of risk factors, and may not easily be summarized in a few measures.

Example: As an example of the interpretation of a three-factor regression model, suppose that in (6.9) the three main effects are present along with the two-factor AC interaction. Assume further that the values of the parameters are given by

$$\exp(\beta_1) = \psi_{100} = 2$$
$$\exp(\beta_2) = \psi_{010} = 3$$
$$\exp(\beta_3) = \psi_{001} = 4$$

and

$$\exp(\gamma_{13}) = \frac{\psi_{101}}{\psi_{100}\psi_{001}} = 2.$$

Then we can reconstruct the seven odds ratios for the three-dimensional cross-classification as the entries in the tables

	Factor C +		Factor C −	
	Factor B		Factor B	
Factor A	+	−	+	−
+	48	16	6	2
−	12	4	3	1

The relative risk of A is twice as great for those exposed to C as for those not so exposed, and vice versa. Otherwise the risks combine in a perfectly multiplicative fashion.

Further details concerning the fitting and interpretation of logistic and log linear models of the type introduced in this section are given in the elementary text by Fienberg (1977). More comprehensive accounts are given by Bishop, Fienberg and Holland (1975), Haberman (1974) and Cox (1970). Vitaliano (1978) conducts an analysis of a case-control study of skin cancer as related to sunlight exposure, using a logistic regression model with four factors, one at four levels and the remainder at two.

6.2 General definition of the logistic model

So far the logistic model has been used solely as a means of relating disease probabilities to one or more categorical risk factors whose levels are represented by indicator variables. More generally the model relates a dichotomous outcome variable y which, in our context, denotes whether (y = 1) or not (y = 0) the individual develops the disease during the study period, to a series of K regression variables $\underset{\sim}{x} = (x_1, \ldots, x_K)$ *via* the equation

$$\text{pr}(y = 1 \mid x) = \frac{\exp(\alpha + \Sigma\beta_k x_k)}{1 + \exp(\alpha + \Sigma\beta_k x_k)} \qquad (6.10)$$

or, equivalently,

$$\text{logit pr}(y = 1 \mid x) = \alpha + \sum_{k=1}^{K}\beta_k x_k.$$

This formulation implies that the relative risk for individuals having two different sets $\underset{\sim}{x}^*$ and x of risk variables is

$$\text{RR} = \frac{P(x^*)\ \{1-P(x)\}}{P(x)\ \{1-P(x^*)\}} = \exp\{\sum_{k=1}^{K}\beta_k(x_k^* - x_k)\}. \qquad (6.11)$$

Thus α represents the log odds of disease risk for a person with a standard ($\mathbf{x} = \mathbf{O}$) set of regression variables, while $\exp(\beta_k)$ is the fraction by which this risk is increased (or decreased) for every unit change in x_k. A large number of possible relationships may be represented in this form by including among the x's indicator variables and continuous measurements, transformations of such measurements, and cross-product or interaction variables.

As we saw in the last chapter, one important means of controlling the effects of nuisance or confounding variables is by stratification of the study population on the basis of combinations of levels of these variables. When conducting similar analyses in the context of logistic regression, it is convenient to generalize the model further so as to isolate the stratum effects, which are often of little intrinsic interest, from the effects of the risk factors under study. With $P_i(\mathbf{x})$ denoting the disease probability in stratum i for an individual with risk variables $\mathbf{x}$, we may write

$$\text{logit } P_i(\mathbf{x}) = \alpha_i + \sum_{k=1}^{K} \beta_k x_k. \tag{6.12}$$

If none of the regression variables are interaction terms involving the factors used for stratification, a consequence of (6.12) is that the relative risks associated with the risk factors under study are constant over strata. By including such interaction terms among the x's, one may model changes in the relative risk which accompany changes in the stratification variables. The fact that the parameters of the logistic model are so easily interpretable in terms of relative risk is, as we have said, one of the main reasons for using the model.

The earliest applications of this model were in prospective studies of coronary heart disease in which $\mathbf{x}$ represented such risk factors as age, blood pressure, serum cholesterol and cigarette consumption (Cornfield, 1962; Truett, Cornfield & Kannel, 1967). In these investigations the authors used linear discriminant analysis to estimate the parameters, an approach which is strictly valid only if the x's have multivariate normal distributions among both diseased and non-diseased (see § 6.3). The generality of the method was enhanced considerably by the introduction of maximum likelihood estimation procedures (Walker & Duncan, 1967; Day & Kerridge, 1967; Cox, 1970). These are now available in several computer packages, including the General Linear Interactive Modelling system (GLIM) distributed by the Royal Statistical Society (Baker & Nelder, 1978).

We noted in § 2.8 that for a long study it is appropriate to partition the time axis into several intervals and use these as one of the criteria for forming strata. In the present context this means that the quantity $P_i(\mathbf{x})$ refers more specifically to the *conditional* probability of developing disease during the time interval specified by the i^{th} stratum, given that the subject was disease-free at its start. For follow-up or cohort studies, if we are to use conventional computer programmes for logistic regression with conditional probabilities, separate data records must be read into the computer for *each stratum* in which an individual appears. Thomson (1977) discusses in some detail the problems of estimation in this situation.

A limiting form of the logistic model for conditional probabilities, obtained by allowing the time intervals used for stratification to become infinitesimally small, is known as the *proportional hazards* model (Cox, 1972). Here the ratio of incidence rates for individuals with exposures $\mathbf{x}^*$ and $\mathbf{x}$ is given exactly by the right-hand side of equation (6.11). This approach has the conceptual advantage of eliminating the odds ratio approximation altogether, and thus obviates the rare disease assumption. The model has a history of successful use in the statistical analysis of survival studies, and it is becoming increasingly clear that many of the analytic techniques developed for use in that field can also be applied in epidemiology (Breslow, 1975, 1978). Prentice and Breslow (1978) present a detailed mathematical treatment of the role of the proportional hazards model in the analysis of case-control study data. Methodological techniques stemming from the model are identical to those presented in Chapter 7 on matched data.

6.3 Adaptation of the logistic model to case-control studies[1]

According to the logistic model as just defined, the exposures $\mathbf{x}$ are regarded as fixed quantities while the response variable y is random. This fits precisely the cohort study situation because it is not known in advance whether or not, or when, a given individual will develop the disease. With the case-control approach, on the other hand, subjects are selected on the basis of their disease status. It is their history of risk factor exposures, as determined by retrospective interview or other means, which should properly be regarded as the random outcome. Thus an important question, addressed in this section, is: how can the logistic model for disease probabilities, which has such a simple and desirable interpretation *vis-à-vis* relative risk, be adapted for use with a sample of cases and controls?

If there is but a single binary risk factor with study subjects classified simply as exposed *versus* unexposed, the answer to this is perfectly clear. Recall first of all our demonstration in § 2.8 that the odds ratio ψ of disease probabilities for exposed *versus* unexposed is identical to the odds ratio of exposure probabilities for diseased *versus* disease-free. When drawing inferences about ψ on the basis of data in 2×2 tables (4.1), it makes absolutely no difference whether the marginal totals m_1 and m_0 corresponding to the two exposure categories are fixed, as in a cohort study, or whether the margins n_1 and n_0 of diseased and disease-free are fixed, as in a case-control study. The estimates, tests and confidence intervals for ψ derived in § 4.1 and § 4.2 in no way depend on how the data in the tables are obtained. Hence we have already demonstrated for 2×2 tables that *inferences about relative risk are made by applying to case-control data precisely the same set of calculations as would be applied to cohort data from the same population.*

This identity of inferential procedures, whether sampling is carried out according to a cohort or case-control design, is in fact a fundamental property of the general logistic model. We illustrate this feature with a simple calculation involving conditional probabilities (Mantel, 1973; Seigel & Greenhouse, 1973) which lends a good deal of plausibility to the deeper mathematical results discussed afterwards. It suffices to consider the model (6.10) for disease probabilities in a single population, as results for the

[1] This section, which is particularly abstract, deals with the logical basis for the application of logistic regression to case-control data. Readers interested only in practical applications can go directly to § 6.5.

stratified situation are quite analogous. Suppose the indicator variable z denotes whether (z = 1) or not (z = 0) someone is sampled, and let us define

$$\pi_1 = \text{pr}(z = 1 \mid y = 1)$$

to be the probability that a diseased person is included in the study as a case and

$$\pi_0 = \text{pr}(z = 1 \mid y = 0)$$

to be the probability of including a disease-free person in the study as a control. Typically π_1 is near unity, i.e., most potential cases are sampled for the study, while π_0 has a lower order of magnitude.

Consider now the conditional probability that a person is diseased, given that he has risk variables **x** *and that he was sampled for the case-control study.* Using Bayes' Theorem (Armitage, 1975) we compute pr(y = 1 | z = 1,**x**)

$$= \frac{\text{pr}(z = 1 \mid y = 1,\mathbf{x})\,\text{pr}(y = 1 \mid \mathbf{x})}{\text{pr}(z = 1 \mid y = 0,\mathbf{x})\,\text{pr}(y = 0 \mid \mathbf{x}) + \text{pr}(z = 1 \mid y = 1,\mathbf{x})\,\text{pr}(y = 1 \mid \mathbf{x})}$$

$$= \frac{\pi_1 \exp(\alpha + \Sigma\beta_k x_k)}{\pi_0 + \pi_1 \exp(\alpha + \Sigma\beta_k x_k)}$$

$$= \frac{\exp(\alpha^* + \Sigma\beta_k x_k)}{1 + \exp(\alpha^* + \Sigma\beta_k x_k)}$$

where $\alpha^* = \alpha + \log(\pi_1/\pi_0)$. In other words, the disease probabilities for those in the sample continue to be given by the logistic model with precisely the same βs, albeit a different value for α. This observation alone would suffice to justify the application of (6.10) to case-control data provided we could also assume that the probabilities of inclusion in the study were independent for different individuals. However, unless a separate decision was made on whether or not to include each potential case or control in the sample, this will not be true. In most studies some slight dependencies are introduced because the total numbers of cases and controls are fixed in advance by design. Hence a somewhat more complicated theory is required.

One assumption made implicitly in the course of this derivation deserves further emphasis. This is that *the sampling probabilities depend only on disease status and not on the exposures.* In symbols, $\text{pr}(z = 1 \mid y,\mathbf{x}) = \text{pr}(z = 1 \mid y) = \pi_y$ for y = 1 and 0. With a stratified design and analysis these sampling fractions may vary from stratum to stratum, but again should not depend on the values of the risk variables. An illustration of the magnitude of the bias which may accompany violations of this assumption was made earlier in § 2.8.

Since case-control studies typically involve *separate samples* of fixed size from the diseased and disease-free populations, the independent probabilities are those of risk variables given disease status. If the sample contains n_1 cases and n_0 controls, the likelihood of the data is a product of n_1 terms of the form $\text{pr}(\mathbf{x} \mid y = 1)$ and n_0 of the form $\text{pr}(\mathbf{x} \mid y = 0)$. Using basic rules of conditional probability, each of these can be expressed

$$\mathrm{pr}(\mathbf{x}|\mathrm{y}) = \frac{\mathrm{pr}(\mathrm{y}|\mathbf{x})\mathrm{pr}(\mathbf{x})}{\mathrm{pr}(\mathrm{y})} \tag{6.13}$$

as the product of the conditional probabilities of disease given exposure, specified by the logistic model, times the ratio of unconditional probabilities for exposure and disease.

How one approaches the estimation of the relative risk parameters β from (6.13) depends to a large extent on assumptions made about the mechanism generating the data, i.e., about the joint probability distribution for **x** and y. The key issue is whether the **x** variables themselves, without knowledge of the associated y's, contain any information about the parameters of interest. Such a condition would be expressed mathematically through dependence of the marginal distribution pr(**x**) on β as well as on other parameters, in which case better estimates of β could in principle be obtained by using the entire likelihood (6.13) rather than by using only the portion of that likelihood specified by (6.10).

An example in which the **x**'s do contain information on their own about the relative risk was alluded to in § 6.2. In early applications of logistic regression to cohort studies, the regression variables were assumed to have multivariate normal distributions in each disease category (Truett, Cornfield & Karrel, 1967). If such distributions are centred around expected values of μ_1 for diseased individuals and μ_0 for controls, and have a common covariance matrix Σ, then the corresponding relative risk parameters can be computed to be

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0).$$

Estimation of β from the full likelihood (6.13) thus entails calculation of the sample means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_0$ of the regression variables among cases and controls, of the pooled covariance matrix S_p^2, and substitution of these quantities in place of μ_1, μ_0 and Σ, respectively. While this procedure yields the most efficient estimates of β *provided the assumptions of multivariate normality hold,* severe bias can result if they do not (Halperin, Blackwelder & Verter, 1971; Efron, 1975; Press & Wilson, 1978). It is therefore not recommended for estimation of relative risks, although it may be useful in the early exploratory phases of an analysis to help determine which risk factors contribute significantly to the multivariate equation.

In most practical situations, the **x** variables are distinctly non-normal. Indeed, many if not all of them will be discrete and limited to a few possible values. It is therefore prudent to make as few assumptions as possible about their distribution. This can be accomplished by allowing pr(**x**) in (6.13) to remain completely arbitrary, or else to assume that it depends on a (rather large) set of parameters which are functionally independent of β. Then, following general principles of statistical inference, one could either try to estimate β and pr(**x**) jointly using (6.13); or else one could try to *eliminate* the pr(**x**) term by deriving an appropriate *conditional likelihood* (Cox & Hinkley, 1974).

If we decide on the first course, namely joint estimation, a rather remarkable thing happens. Providing pr(**x**) is assumed to remain completely arbitrary, the joint maximum likelihood estimate $\hat{\beta}$ turns out to be identical to that based only on the portion of the likelhood which is specified by the linear logistic model. Furthermore, the standard errors and covariances for $\boldsymbol{\beta}$ generated from partial and full likelihoods also agree. This fact was first noted by Anderson (1972) for the case in which $\mathbf{x}$ was a discrete variable, and established for the general situation by Prentice and Pyke (1979).

Another approach to the likelihood (6.13) is to eliminate the nuisance parameters through consideration of an appropriate conditional distribution. Suppose that a case-control study of $n=n_1+n_0$ subjects yields the exposure vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, but it is not specified which of them pertain to the cases and which to the controls. The conditional probability that the first n_1 $\mathbf{x}$'s in fact go with the cases, as observed, and the remainder with the controls may be written

$$\frac{\prod_{j=1}^{n_1} \operatorname{pr}(\mathbf{x}_j \mid y=1) \prod_{j=n_1+1}^{n_1+n_0} \operatorname{pr}(\mathbf{x}_j \mid y=0)}{\sum_{\iota} \prod_{j=1}^{n_1} \operatorname{pr}(\mathbf{x}_{\iota_j} \mid y=1) \prod_{j=n_1+1}^{n_1+n_0} \operatorname{pr}(\mathbf{x}_{\iota_j} \mid y=0)} \tag{6.14}$$

where the sum in the denominator is over all the $\binom{n}{n_1}$ ways of dividing the numbers from 1 to n into one group $\{\iota_1, \ldots, \iota_{n_1}\}$ of size n_1 and its complement $\{\iota_{n_1+1}, \ldots, \iota_n\}$. Using (6.10) and (6.13) it can be calculated that (6.14) reduces to

$$\frac{\prod_{j=1}^{n_1} \exp(\Sigma \beta_k x_{jk})}{\sum_{\iota} \prod_{j=1}^{n_1} \exp(\Sigma \beta_k x_{\iota_j k})}, \tag{6.15}$$

where x_{jk} denotes the value of the kth regression variable for the jth subject and the sum in the denominator is again over all possible choices of n_1 subjects out of n (Prentice & Breslow, 1978; Breslow et al., 1978). This likelihood depends only on the $\boldsymbol{\beta}$ parameters of interest. However, when n_1 and n_0 are large, the number of summands in the denominator is so great as to rule out its use in practice. Fortunately, as these quantities increase, the conditional maximum likelihood estimate and the standard errors based on (6.15) are almost certain to be numerically close to those obtained by applying the unconditional likelihood (6.10) (Efron, 1975; Farewell, 1979).

In summary, unless the marginal distribution of the risk variables in the sample is assumed to contain some information about the relative risk, methods of estimation based on the joint exposure likelihood yield essentially the same numerical results as do those based on the disease probability model. This justifies the application to case-control data of precisely the same analytic techniques used with cohort studies.

6.4 Likelihood inference: an outline[1]

We have now introduced the logistic regression model as a natural generalization of the odds ratio approach to relative risk estimation, and argued that it may be directly

[1] This section also treats material which is quite technical and is not required for appreciation of the applications of the methods. The reader who lacks formal mathematical or statistical training is advised to skim through it on a first reading, and then refer back to the section while working through the examples.

applied to case-control study data with disease status (case *versus* control) treated as the "dependent" or response variable. Subsequent sections of this chapter will illustrate its application to several problems of varying complexity. With one exception, the illustrative analyses may all be carried out using standard computer programmes for the fitting of linear logistic models by maximum likelihood.

Input to GLIM or other standard programmes is in the form of a *rectangular data array,* consisting of a list of values on a fixed number of variables for each subject in the study, with different subjects on different rows. The variables are typically in the order $(y,x_1, ..., x_K)$, where y equals 1 or 0 according to whether the subject is a case or control, while the x's represent various discrete and/or continuous regression variables to be related to y. *Output* usually consists of estimates of the regression coefficients for each variable, a variance/covariance matrix for the estimated coefficients, and one or more test statistics which measure the goodness of fit of the model to the observed data. It is not necessary to have a detailed understanding of the arithmetical operations linking the inputs to the outputs in order to be able to use the programme. Researchers in many fields have long used similar programmes for ordinary (least squares) fitting of multiple regression equations, with considerable success. Nevertheless, some appreciation of the fundamental concepts involved can help to dispel the uneasiness which accompanies what otherwise might seem a rather "black box" approach to data analysis. In this section we outline briefly the key features of likelihood inference in the hopes that it may lay the logical foundation for the interpretation of the outputs. More detailed expositions of this material can be found in the books by Cox (1970), Haberman (1974), Bishop, Fienberg and Holland (1975) and Fienberg (1977).

Statistical inference starts with an expression for the probability, or likelihood, of the observed data. This depends on a number of unknown parameters which represent quantitative features of the population from which the data are sampled. In our situation the likelihood is composed of a product of terms of the form (6.10), one for each subject. The α's and β's are the unknown parameters, interest being focused on the β's because of their ready interpretation *vis-à-vis* relative risk.

Estimates of the parameters are selected to be those values which maximize the likelihood or rather, and what is equivalent, those which maximize its logarithm. The parameters thus estimated, which are often denoted $\hat{\alpha}$ and $\hat{\beta}$, are inserted back into the individual likelihoods (6.10) to calculate the *fitted* or *predicted* probability $\hat{P}$ of being a case for each study subject. If we subtract twice the maximized log likelihood from zero, which is the absolute maximum achieved as all the fitted values $\hat{P}$ approach the observed y's, and sum up over all individuals in the sample, we obtain the expression

$$G = -2\Sigma\{y\log \hat{P} + (1-y)\log(1-\hat{P})\} \tag{6.16}$$

for the *log likelihood statistic*[1]. Although G as given here does not have any well defined distribution itself, differences between G statistics for different models may be interpreted as chi-squares (see below).

Other important statistics in likelihood analysis are defined in terms of the first and second derivatives of the log likelihood function. The vector of its first partial derivatives

[1] The statistic (6.16) is called the *deviance* in GLIM.

is known as the efficient score, $\mathbf{S} = \mathbf{S}(\alpha,\beta)$, while the negative of the matrix of second partial derivatives is the *information matrix,* denoted $\mathbf{I} = \mathbf{I}(\alpha,\beta)$. The variance/covariance matrix of the estimated parameters is obtained from the inverted information matrix, evaluated at the maximum likelihood estimate (MLE):

$$\text{Covariance matrix for } (\hat{\alpha},\hat{\beta}) = \mathbf{I}^{-1}(\hat{\alpha},\hat{\beta}). \tag{6.17}$$

Another specification of the MLE is as the value α,β for which the efficient score is zero.

Likelihood inference typically proceeds by fitting a *nested hierarchy* of models, each one containing the last. For example, we might start with the model

$$(1) \qquad \text{logit pr}(y \mid \mathbf{x}) = \alpha$$

which specifies that the disease probabilities do not depend on the regression variables, i.e., that the log relative risk for different x's is zero. This would be elaborated in a second model

$$(2) \qquad \text{logit pr}(y \mid \mathbf{x}) = \alpha + \beta_1 x_1,$$

for which the log relative risk associated with risk factor x_1 is allowed to be non-zero. A further generalization is then to

$$(3) \qquad \text{logit pr}(y \mid \mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

in which the coefficients for two more variables, one of which might for instance be an interaction involving x_1, are also allowed to be non-zero.

At each stage we obtain the MLEs of the coefficients in the model, together with their estimated variances and covariances. We also carry out a test for the significance of the additional parameters, which is logically equivalent to testing whether the current model fits better than the last one. Three tests are available. The *likelihood ratio test* is simply the difference of the maximized log likelihood statistics (6.16) for the two models. If G_1, G_2 and G_3 denote the values of these statistics for models 1, 2 and 3, respectively, then necessarily $G_3 \leqq G_2 \leqq G_1$. Each hypothesis is less restrictive than the last and its fitted probabilities $\hat{P}$ will therefore generally be closer to the observed y's. $G_1 - G_2$ tests the hypothesis $\beta_1 = 0$, i.e., the significance of x_1 as a risk factor, while G_2–G_3 evaluates the additional contributions of x_2 and x_3 after the effect of x_1 is accounted for.

The *score statistic* for testing the significance of the additional parameters is based on the efficient score evaluated at the MLE for the previous model, appropriately augmented with zeros. For example, the score test of Model 2 against Model 1 is given by

$$S_2 = \mathbf{S}(\hat{\alpha},0)^{\mathrm{T}}\, \mathbf{I}^{-1}(\hat{\alpha},0)\, \mathbf{S}(\hat{\alpha},0) \tag{6.18}$$

where $\mathbf{S}$ and $\mathbf{I}$ are calculated for Model 2 whereas $\hat{\alpha}$ is the MLE for Model 1. Similarly the score test of the hypothesis $\beta_2 = \beta_3 = 0$ in Model 3 is

$$S_3 = \mathbf{S}(\hat{\alpha},\hat{\beta}_1,0,0)^{\mathrm{T}}\, \mathbf{I}^{-1}(\hat{\alpha},\hat{\beta}_1,0,0)\, \mathbf{S}(\hat{\alpha},\hat{\beta}_1,0,0).$$

A third test for the significance of the additional parameters in a model is simply to compare their estimated values against 0, using their standard errors as a reference.

Thus to test $\hat{\beta}_1 = 0$ in Model 2 we would calculate the *standardized regression coefficient*

$$Z_1 = \frac{\hat{\beta}_1}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}}$$

where $\mathrm{Var}(\hat{\beta}_1)$ was the appropriate diagonal term in the inverse information matrix for Model 2. A test statistic analogous to the previous two is based on the square of this value

$$Z_1^2 = \frac{\hat{\beta}_1^2}{\mathrm{Var}(\hat{\beta}_1)}.$$

Similarly, the test of $\beta_2 = \beta_3 = 0$ in Model 3 is given by the statistic

$$(\hat{\beta}_2,\hat{\beta}_3)\,\Sigma_{23}^{-1}\begin{pmatrix}\hat{\beta}_2\\ \hat{\beta}_3\end{pmatrix},$$

where Σ_{23} is the estimated variance/covariance matrix for $(\hat{\beta}_2,\hat{\beta}_3)$ in Model 3.

In large samples all three of these statistics are known to give approximately equal numerical results under the null hypothesis, and to have distributions which are chi-square with degrees of freedom equal to the number of *additional* parameters (Rao, 1965). In other words, if Model 1 holds we should have approximately

$$G_1 - G_2 \approx S_2 \approx Z_1^2 \approx \chi_1^2. \tag{6.19}$$

Similarly, all three statistics for the hypothesis $\beta_2 = \beta_3 = 0$ in Model 3 should yield similar numerical results, and will have approximate χ_2^2 distributions, if Model 2 adequately summarizes the data. The first and third statistics are most easily calculated from the output of standard programmes such as GLIM. The score statistic, while not routinely calculated by standard programmes, is mentioned here for two reasons. First, in simple situations it is identical with the elementary test statistics presented in Chapter 4, and thus provides a link between the two approaches (Day & Byar, 1979). Second, the nominal chi-square distribution is known to approximate that of the score statistic more closely in small samples, so that its use is less likely to lead to erroneous conclusions of statistical significance (Lininger et al., 1979).

Two other statistics should be mentioned which are useful for evaluating goodness of fit with *grouped data.* These arise when there are a limited number of distinct risk categories, i.e., when the number of **x** values is sufficiently small compared with the size of the study population that quite a few individuals within each stratum have the same **x**. In this case, rather than consider each data record on its own for the analysis, it makes sense to group together those records within each stratum which have the same set of exposures. Suppose that N denotes the total number of individuals in a particular group, of whom n_1 are cases and n_0 are controls. Since the exposures are identical, the estimated probabilities $\hat{P}$ will apply equally to everyone in the group. $N\hat{P}$ may therefore be interpreted as the *expected* or fitted number of cases, while $N(1-\hat{P})$ is the expected number of controls. An appropriate version of the likelihood ratio statistic for this situation is

$$G = 2\Sigma[n_1\log(n_1/N\hat{P}) + n_0\log\{n_0/N(1-\hat{P})\}] \qquad (6.20)$$

where the sum is over all the distinct groups or risk categories. Another measure of goodness of fit of model to data is the ubiquitous chi-square statistic

$$\tilde{G} = \Sigma\left[\frac{(n_1-N\hat{P})^2}{N\hat{P}} + \frac{\{n_0-N(1-\hat{P})\}^2}{N(1-\hat{P})}\right]. \qquad (6.21)$$

Unless the data are quite "thin", so that the fitted values of cases or controls for many groups are less than five, these two expressions should yield reasonably close numerical answers when the model holds.

The formulae (6.20) and (6.21) may be expressed in more familiar terms, as functions of the observed (O) and expected (E) numbers in each cell, provided we remember that the cases and controls in each group constitute *separate* cells and thus make separate contributions. The likelihood ratio statistic becomes

$$G = 2\Sigma O \log(O/E), \qquad (6.22)$$

while the chi-square measure is

$$\tilde{G} = \Sigma\frac{(O-E)^2}{E}. \qquad (6.23)$$

Provided the number of groups is small in relation to the total number of cases, each of the statistics G and $\tilde{G}$ have asymptotic chi-square distributions under the null hypothesis. Degrees of freedom are equal to the number of groups less the number of parameters in the logistic model. While they provide us with an overall evaluation of how well the model conforms to the data, these tests may be rather insensitive to particular types of departure from the model. Better tests are obtained by constructing a more general model, with a limited number of additional parameters which express the nature of the departure, and then testing between the two models as outlined earlier.

It should be emphasized that the methods discussed in this section, and illustrated in the remainder of the chapter, are based on *unconditional likelihoods* (6.10) and (6.12) and involve explicit estimation of the α nuisance parameters as well as of the β's. For some of the simpler problems, e.g., the combination of results from 2×2 tables, inference may be carried out also in terms of conditional likelihoods which depend only on the parameters of interest. If the number of nuisance parameters is large, and the data thin, this approach avoids some well known problems of bias (see § 7.1). It also enables exact inferences to be made (§ 4.2). Since many of the procedures in Chapter 4 and all of those in Chapters 5 and 7 are based on such conditional likelihoods, the methods discussed there would be expected to yield more accurate results for finely stratified or matched data than those presented in this chapter. However, the exact conditional procedures are too burdensome computationally for many of the problems which confront us. Thus, while we may lose some accuracy with the logistic regression approach, what we gain in return is a coherent methodology capable of handling a wide variety of problems in a uniform manner.

6.5 Combining results from 2 × 2 tables

As our first worked example using the logistic model, we return to the problem of combining information about the relative risk from a series of 2×2 tables. In this case there is a single exposure variable x, coded x = 1 for exposed and x = 0 for unexposed. The model (6.12) for the probabilities $P_i(x)$ of disease in the i^{th} of I strata becomes

$$\operatorname{logit} P_i(x) = \alpha_i + \beta x,$$

which expresses the idea that the relative risks in each stratum are given by the constant $\psi = \exp(\beta)$.

Simultaneous estimation of the α_i and β parameters as outlined in the last section leads to the estimate $\hat{\psi} = \exp(\hat{\beta})$ identified in § 4.4 as the unconditional or asymptotic maximum likelihood estimate (MLE). This has the property that the sum of the fitted values of exposed cases over all I strata is equal to the sum of the observed values. More precisely, suppose the data are laid out as in (4.21). Denote the fitted values by

$$\begin{aligned} \hat{a}_i &= m_{1i}\hat{P}_{1i} = \frac{m_{1i}\exp(\hat{\alpha}_i+\hat{\beta})}{1+\exp(\hat{\alpha}_i+\hat{\beta})} \\ \hat{b}_i &= m_{0i}\hat{P}_{0i} \quad \frac{m_{0i}\exp(\hat{\alpha}_i)}{1+\exp(\hat{\alpha}_i)} \end{aligned} \qquad (6.24)$$

and for the remaining cells by subtraction, $\hat{c}_i = m_{1i} - \hat{a}_i$, $\hat{d}_i = m_{0i} - \hat{b}_i$. Agreement of the observed and marginal totals means $\Sigma\hat{a}_i = \Sigma a_i$, $\Sigma\hat{b}_i = \Sigma b_i$, and so on. Since the squared deviations of observed and fitted values for the four cells in each stratum agree, i.e., $(a_i-\hat{a}_i)^2 = (b_i-\hat{b}_i)^2 = (c_i-\hat{c}_i)^2 = (d_i-\hat{d}_i)^2$, it follows that the chi-square statistic (6.23) for testing goodness of fit of the model may be written

$$\tilde{G} = \sum_{i=1}^{I} (a_i - \hat{a}_i)^2 \left\{ \frac{1}{\hat{a}_i} + \frac{1}{\hat{b}_i} + \frac{1}{\hat{c}_1} + \frac{1}{\hat{d}_i} \right\} = \sum_{i=1}^{I}(a_i - \hat{a}_i)^2/\mathrm{Var}(a_i),$$

where we have used the variance formula (4.13). This chi-square agrees precisely with the goodness of fit statistic (4.30) derived earlier, except that we now use MLE for the parameters.

Example: To illustrate these calculations we reanalyse the grouped data from the Ille-et-Vilaine study of oesophageal cancer summarized in Table 4.1. Here the six strata are defined as ten-year age groups from 25–34 through 75+ years, while average alcohol consumption is treated as a binary risk factor with 0–79 g/day (up to one litre of wine) representing "unexposed" and anything over this amount "exposed". The data would be rearranged for computer entry as shown in Table 6.1, where 12 risk categories or groups are defined by the six strata and two levels of exposure. Within each of these the total N of cases + controls is regarded as the denominator of an observed disease proportion, while the number of cases is the numerator. The numerical results should be compared closely with those already obtained in § 4.4.

Table 6.1 Data from Table 4.5 reorganized for entry into a computer programme for linear logistic regression

Age stratum	Exposure (x=1 for 80+ g/day)	Cases	Total (cases + controls)
1	1	1	10
1	0	0	106
2	1	4	30
2	0	5	169
3	1	25	54
3	0	21	159
4	1	42	69
4	0	34	173
5	1	19	37
5	0	36	124
6	1	5	5
6	0	8	39

Results of fitting several versions of the model to these data are summarized in Table 6.2. In the first version, with six parameters, the disease probabilities may vary with each age group but not with exposure ($\beta = 0$). Considering the huge goodness of fit statistics, this assumption is clearly not tenable ($p<0.00001$). When a single relative risk parameter (β) is introduced the fit improves considerably. However, the chi-square ($\tilde{G} = 9.32$, $p = 0.15$) and log likelihood ($G = 11.04$, $p = 0.05$) statistics give somewhat different answers as to whether the differences in relative risk between strata are significantly different. Both are sufficiently large to alert us to the possibility of systematic variations in the relative risk for different age groups, which should be investigated further.

Inferences about the relative risk are made in terms of the estimate $\hat{\beta} = 1.670$ and its *standard error* 0.190[1]. We compute $\hat{\psi} = \exp(1.670) = 5.31$ as the point estimate of relative risk. Ninety-five percent confidence limits for β are given by $\beta_L = 1.670-1.96 \times 0.190 = 1.30$ and $\beta_U = 1.670+1.96 \times 0.190 = 2.04$. These correspond to bounds of $\psi_L = \exp(\beta_L) = 3.66$ and $\psi_U = \exp(\beta_U) = 7.71$ on the relative risk, which compare well with those derived in § 4.4 using two other methods.

The third model shown in Table 6.2 was fitted to see whether there was a *systematic trend in relative risk with age.* This took the form

$$\text{logit } P_i(x) = \alpha_i + \beta x + \gamma x(i-3.5),$$

where now β represents the log relative risk for a "typical" age ($i = 3.5$), while γ represents the linear trend in this depending on the age group indicator i. The lack of a significant improvement in the goodness of fit statistics, and the small value of $\hat{\gamma}$ as compared with its standard error, tell us that there is little evidence for such a trend.

More information about the sources of departure from model assumptions can be obtained from an examination of the *residuals,* the differences between the observed and fitted numbers of disease cases in each category (Table 6.3). As an illustration of their calculation, the fitted values for the 35–44 age category are found from (6.24) and the estimated coefficients in Table 6.2 to be

$$\hat{a}_2 = \frac{30 \times \exp(-3.512+1.670)}{1+\exp(-3.512+1.670)} = 4.10$$

and

$$\hat{b}_2 = \frac{169 \times \exp(-3.512)}{1+\exp(-3.512)} = 4.90.$$

[1] The standard error of an estimate is the square root of its estimated variance.

Table 6.2 Results of fitting several versions of the linear logistic model (6.3) to the data in Table 6.1

Model	No. of parameters	DF	Goodness-of-fit statistics		Regression coefficients Age strata (years)						Log relative risk and interactions	
			Log likelihood G	Chi-square $\tilde{G}$	25–34 $\hat{\alpha}_1$	35–44 $\hat{\alpha}_2$	45–54 $\hat{\alpha}_3$	55–64 $\hat{\alpha}_4$	65–74 $\hat{\alpha}_5$	75+ $\hat{\alpha}_6$	Alcohol $\hat{\beta}$ ± S.E.	Alcohol × age $\hat{\gamma}$ ± S.E.
1	6	6	90.56	101.80	–4.746	–3.050	–1.289	–0.781	–0.656	–0.869	–	–
2	7	5	11.04	9.32	–5.054	–3.512	–1.855	–1.341	–1.087	–1.092	1.670 ± 0.190	–
3	8	4	10.61	8.48	–5.182	–3.617	–1.900	–1.334	–1.049	–1.055	1.714 ± 0.201	0.125 ± 0.189

Table 6.3 Residuals from fitting model 2 of Table 6.2 to data in Table 6.1

Age stratum (years)	Exposure	Numbers of cases Observed	Expected	Variance	Standardized residual
25–34	1	1	0.33	0.32	1.19
	0	0	0.67	0.67	–0.82
35–44	1	4	4.10	3.54	–0.06
	0	5	4.90	4.75	0.05
45–54	1	25	24.50	13.4	0.14
	0	21	21.50	18.6	–0.12
55–64	1	42	40.13	16.8	0.46
	0	34	35.87	28.4	–0.35
65–74	1	19	23.74	8.51	–1.63
	0	36	31.26	23.4	0.98
75+	1	5	3.20	1.15	1.68
	0	8	9.80	7.34	–0.66
Total		200	200.00		

We easily verify that the sum of the fitted numbers of exposed cases over the six strata, 0.33 + 4.10 + 24.50 + 40.13 + 23.74 + 3.20, equals the sum of the observed number, namely 96. This confirms the property of the maximum likelihood estimate mentioned earlier.

Variances for the O–E residuals are calculated as $N\hat{P}\hat{Q}$, where N is the denominator (total of cases and controls), $\hat{P}$ is the estimated disease probability and $\hat{Q} = 1 - \hat{P}$. Dividing each residual by its standard error gives us the *standardized residuals,* which when squared and summed produce the $\tilde{G}$ goodness of fit statistic. The greatest contribution to this comes from the last two age groups. For the 65–74 year-olds the deficit of 19 exposed cases compared with 23.74 expected indicates a relative risk smaller than that of the other groups; while for the 75+ group the excess of 5 observed to 3.20 expected implies a larger than average relative risk. The contribution from the youngest age group can be largely discounted because only one case appears. Since there does not seem to be any obvious pattern to the residuals, we feel comfortable in attributing the observed departures from the fitted model to chance phenomena.

6.6 Qualitative analysis of grouped data from Ille-et-Vilaine

In § 4.6 we applied classic Mantel-Haenszel methodology to study the joint effects of two risk factors, alcohol and tobacco, on the relative risk of oesophageal cancer in Ille-et-Vilaine. Both factors were partitioned into four levels, yielding 16 risk categories in all. Our first approach was to compute separate estimates of the age-adjusted relative risk for each such category, assigning the value 1.0 to the low alcohol, low tobacco cell. Later we estimated relative risks for each alcohol level, simultaneously adjusting for age and tobacco, and each tobacco level, simultaneously adjusting for alcohol and age. This was a cumbersome procedure which required that we construct and summarize several different series of 24 2×2 tables. The relative risks obtained for each alcohol and tobacco level were multiplied together to estimate the joint effect of these two variables. However, there was no very satisfactory way of testing the validity of the multiplicative hypothesis, and the relative risks obtained in this fashion lacked the desirable property of consistency.

In this section we demonstrate that a comprehensive and integrated analysis, which parallels the Mantel–Haenszel approach, may be carried out quite simply using the logistic model with stratification (6.12). The starting point is the grouping of the 200 cases and 775 controls into $4 \times 4 \times 6 = 96$ cells, each of which represents a combination of the categories of alcohol, tobacco and age. According to the principles outlined in § 6.3, the observations in each cell are treated in the statistical analysis as independent binomial observations, with cases representing the numerator and cases + controls the denominator. Appendix I lists the 96 binomial observations so formed. In fact, since 8 of the cells were devoid of cases and controls there are effectively only 88 observations and it is this figure that one uses to determine degrees of freedom.

As only the qualitative or categorical aspects of the data are to be considered here, the regression variables x appearing in the model are indicator variables which take the value 1 or 0 according to whether the cell (observation) in question corresponds to a given level or combination of levels of the various study factors. Even the parameter α_i in (6.12) can be regarded as the coefficient of an indicator variable which takes the value 1 for the i^{th} stratum and 0 otherwise. Sophisticated programmes such as GLIM will automatically construct such indicators for all factors specified by the user as being categorical.

Table 6.4 shows explicitly the values of the regression variables so constructed. Since they depend only on alcohol and tobacco it suffices to show their values for the first age group only. The first three variables define the main effects of each alcohol category on risk, while the next three define the main effects of tobacco. Thus, $x_2 = 1$ for the third alcohol group and 0 otherwise, while $x_6 = 1$ for the fourth tobacco group and 0 otherwise. Cells having 0 values for all six of these variables correspond to the lowest consumption levels of both factors and are assigned a baseline relative risk of unity.

Variables x_7 to x_{15} define the totality of qualitative interactions between alcohol and tobacco. They are obtained by multiplying together the dummy variables representing the main effects:

$$x_7 = x_1 \times x_4;\ x_8 = x_1 \times x_5;\ \ldots;\ x_{15} = x_3 \times x_6.$$

Inclusion of all six main effect and all nine interaction variables in the equation imposes no constraints on how the relative risks vary over the 16 alcohol/tobacco cells. The 15 parameters in the model yield 15 estimated relative risks, with the value 1.0 being assigned to the baseline category. Thus the log relative risk for the third alcohol and fourth tobacco group is estimated as $\hat{\beta}_2 + \hat{\beta}_6 + \hat{\beta}_{12}$, i.e., as a contribution from the alcohol level plus one from the tobacco level plus the interaction. One obvious drawback to this method of parameterizing the interactions is that it does not lead to the ready identification of quantitative patterns which may be of particular interest. Alternative parameterizations are considered in the next section.

Table 6.5 summarizes the results of fitting several regression models using qualitative regression variables. By subtracting the goodness-of-fit (G) measures for Models 2 and 3 from that for Model 1 we obtain χ_3^2 statistics of 141.0 and 36.6, respectively, for testing the significance of alcohol and tobacco, *ignoring* the effects of the other variable. Both factors have an enormous influence on risk. Subtracting the G's for Model 4 from those for Models 2 and 3 yields χ_3^2 statistics of 128.0 and 23.6. These determine the significance of alcohol and tobacco while *adjusting* for the effects of the other variable. The adjusted chi-squares are a little smaller than the unadjusted ones, reflecting the slight correlation between alcohol and tobacco consumption. However, their magnitude

Table 6.4 Values of qualitative risk variables for the first 16 of 96 grouped data records: Ille-et-Vilaine study of oesophageal cancer

Obser-	Levels of			Alcohol main			Tobacco main			Alcohol × tobacco interaction								
vation	age	alc	tob	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
3	1	1	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
4	1	1	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
5	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	2	2	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
7	1	2	3	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0
8	1	2	4	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0
9	1	3	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	3	2	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0
11	1	3	3	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
12	1	3	4	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
13	1	4	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
14	1	4	2	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0
15	1	4	3	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0
16	1	4	4	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1

Table 6.5 Summary of goodness of fit of several logistic regression models: grouped data from the Ille-et-Vilaine study of oesophageal cancer

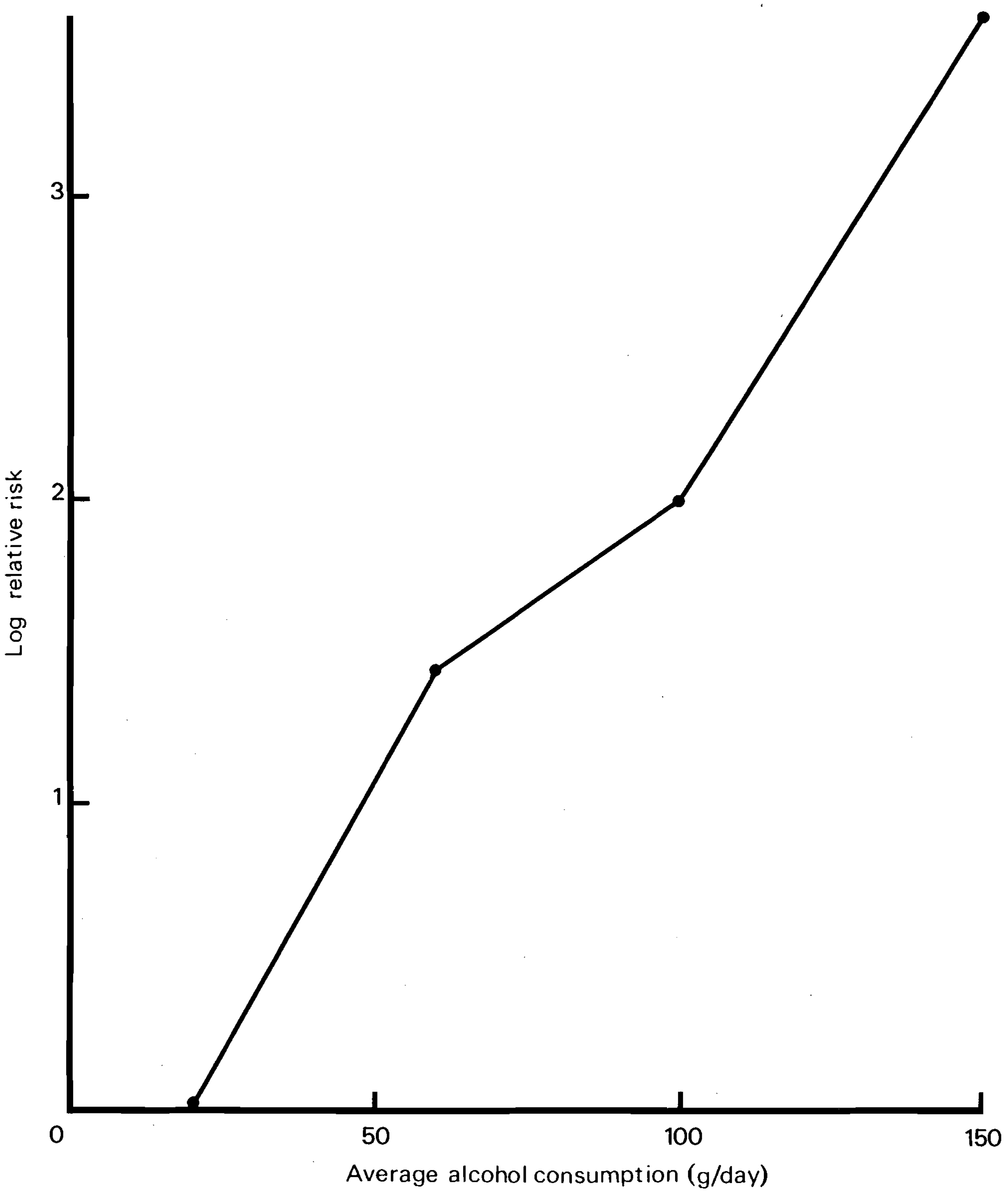
Model	Regression variables included[a]	No. of parameters	DF	Goodness of fit G	Hypothesis tested and/or interpretation
1	Age	6	82	246.9	No effect of alcohol or tobacco
2	Age Alcohol (1–3)	9	79	105.9	Effect of alcohol only, adjusted for age
3	Age Tobacco (4–6)	9	79	210.3	Effect of tobacco only, adjusted for age
4	Age Alcohol (1–3) Tobacco (4–6)	12	76	82.3	Main effects for alcohol and tobacco (multiplicative hypothesis), adjusted for age

[a] Numbers in parentheses correspond to variable numbers shown in Table 6.4

indicates that both variables have strong *independent* effects which are not explained by the contribution of the other.

The estimated regression coefficients for Model 2, when exponentiated, yield estimates of the risk for each alcohol level relative to baseline (0–39 g/day) which are adjusted for age but not for tobacco. Thus, $\exp(\hat{\beta}_1) = \exp(1.43) = 4.2$ is the relative risk for the 40–79 g/day group, while for the higher levels of consumption the figures are $\exp(\hat{\beta}_2) = 7.4$ and $\exp(\hat{\beta}_3) = 39.7$. These may be contrasted with the correspond-

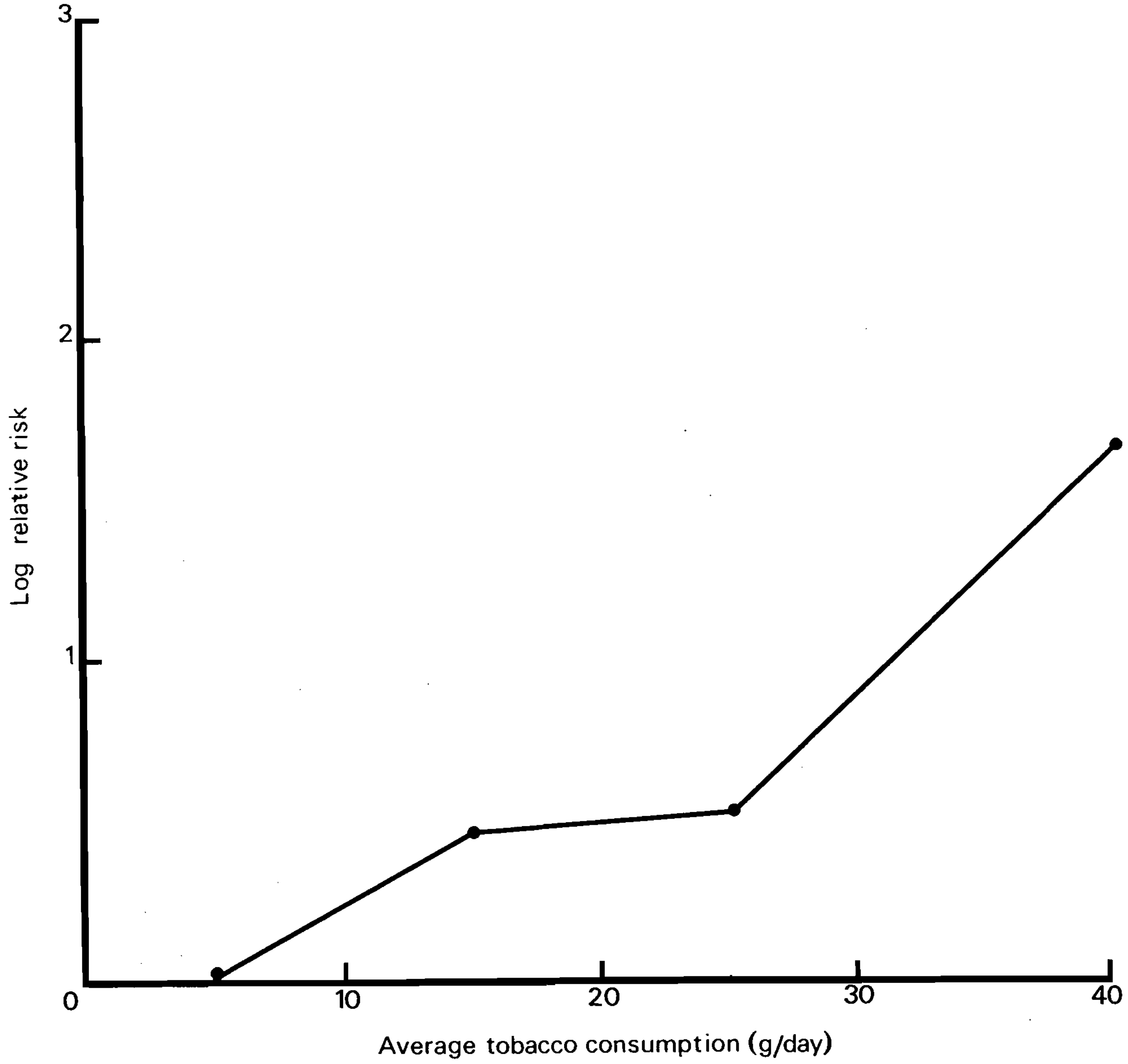
Fig. 6.1 Log relative risk of oesophageal cancer according to four levels of alcohol consumption



ing figures of 4.3, 8.0 and 28.6 obtained by the Mantel-Haenszel (M-H) method (Table 4.4). There is reasonably good agreement except for the highest exposure category, where there were few cases and controls in some strata. For this category the *conditional* maximum likelihood estimate (§ 4.4) was 34.9, midway between the M-H estimate and unconditional MLE. The latter estimate is probably a bit exaggerated here because of the thin data (§ 7.1).

One disadvantage of the elementary methods was that the relative risks obtained upon varying the choice of baseline category were not consistent. For example, the direct M-H estimate of the risk for the fourth alcohol level relative to the second is 8.7 rather than 28.6/4.3 = 6.7. Use of the logistic modelling approach avoids such

Fig. 6.2 Log relative risk of oesophageal cancer according to four levels of tobacco consumption



discrepancies. Estimates of the log relative risks between any two categories are always obtained as the differences in the regression coefficients for those categories (with the proviso that the coefficient for the baseline category is 0), and such differences are not affected by the choice of the baseline category. Thus $\exp(\hat{\beta}_3 - \hat{\beta}_1) = 9.4$ represents the risk of level four relative to level two regardless of how the indicator variables representing the alcohol effects are coded.

Model 4 is the first reasonably satisfactory one in the sense that the goodness of fit chi-square is not significantly higher than its degrees of freedom ($\chi^2_{76} = 82.34$, $p = 0.48$). The fitted regression coefficients are: for alcohol $\hat{\beta}_1 = 1.44$, $\hat{\beta}_2 = 1.98$ and $\hat{\beta}_3 = 3.60$; and for tobacco $\hat{\beta}_4 = 0.44$, $\hat{\beta}_5 = 0.51$ and $\hat{\beta}_6 = 1.64$ (6.11). These show a reasonably smooth linear increase with increasing levels of consumption (Figures 6.1 and 6.2). Taking exponentials, we find estimates of relative risk for each alcohol and tobacco category relative to baseline which, according to the model, combine multiplicatively to yield the results for joint exposures to the two factors shown in Table 6.6. In view of the rather weak correlation between alcohol and tobacco consumption ($\varrho = 0.15$, see Table 4.22), it is not surprising that the alcohol relative risks obtained after adjustment for age and tobacco are only slightly smaller than those obtained after adjustment for age alone. Further evidence for the goodness of fit of the multiplicative model is presented in Table 6.7. Its entries, obtained by summing observed and fitted values over the six age categories, show consistently good agreement throughout the range of both risk factors. The greatest discrepancy is in the baseline category, with nine cases of disease against 13.7 expected. While not statistically significant, the slight lack of fit

Table 6.6 Age-adjusted relative risks for each alcohol/tobacco category according to multiplicative model: Ille-et-Vilaine oesophageal cancer study

Alcohol (g/day)	Tobacco (g/day) 0–9	10–19	20–29	30+
0–39	1.0	1.6	1.7	5.2
40–79	4.2	6.6	7.0	21.8
80–119	7.2	11.3	12.1	37.3
120+	36.6	56.8	61.0	188.7

Table 6.7 Observed and expected (age-adjusted) numbers of cases for each alcohol/ tobacco category according to multiplicative model: Ille-et-Vilaine oesophageal cancer study

Alcohol	Tobacco (g/day)							
(g/day)	0–9		10–19		20–29		30+	
	O	E	O	E	O	E	O	E
0–39	9	13.7	10	7.1	5	3.4	5	4.8
40–79	34	30.0	17	18.9	15	16.0	9	9.1
80–118	19	17.8	19	19.9	6	6.7	7	6.6
120+	16	15.5	12	12.1	7	6.9	10	10.5

for this category indicates that the relative risk for the other levels of exposure might possibly be even greater than that suggested by the model.

One drawback to the choice of grouping intervals used in this analysis is that in neither case does the lowest level correspond to zero consumption. To some extent the choice was dictated by necessity in that no diseased individuals abstained completely from both alcohol and tobacco, and even among controls there were very few who did not consume some alcohol. However there were a substantial number of non-smokers in the population. Thus a similar analysis was carried out using five levels of consumption for each variable: 0–24, 25–49, 50–74, 75–99 and 100+ g/day for alcohol and 0, 1–4, 5–14, 15–29 and 30+ g/day for tobacco. Results shown in Tables 6.8 and 6.9 and in Figures 6.3 and 6.4 confirm the multiplicative relationship and the linear effect of alcohol on the log relative risk. The trend with tobacco, on the other hand, is considerably changed in appearance. Even a small amount appears to increase the risk substantially and there are contra-indications to the linearity of the relationship. Figures 6.3 and 6.4 also show for comparison relative risks estimated from the *quantitative* regression models discussed in the next two sections.

It would be tempting to try to subdivide the alcohol and tobacco variables further, say into ten levels each. However even with five levels per variable there are already 5 × 5 × 6 = 150 groups, and with ten levels there would be 600. The larger the number of parameters in the model, the less information there is available for estimating each one; this is reflected in increased standard errors. Further subdivision would lead one to anticipate increasingly erratic behaviour in the estimates, such as the apparent decrease in risk between the 5–14 and 15–29 g/day tobacco categories (Figure 6.4).

Table 6.8 Estimated relative risks for each alcohol/tobacco category according to the multiplicative model: Ille-et-Vilaine oesophageal cancer study

Alcohol	Tobacco (g/day)				
(g/day)	0	1–4	5–14	15–29	30+
0–24	1.0	4.5	7.2	5.8	19.3
25–49	1.7	7.5	11.9	9.6	31.8
50–74	4.8	21.8	34.8	27.9	92.8
75–99	6.8	30.9	49.4	39.7	131.0
100+	17.5	79.0	126.5	101.5	337.0

Table 6.9 Observed and expected (age-adjusted) numbers of cases for each alcohol/tobacco category according to the multiplicative model: Ille-et-Vilaine oesophageal cancer study

Alcohol	Tobacco (g/day)									
(g/day)	0		1–4		5–14		15–29		30+	
	O	E	O	E	O	E	O	E	O	E
0–24	0	1.3	2	1.7	8	8.2	3	3.1	4	2.7
25–49	1	1.2	6	4.5	12	11.4	5	6.1	4	4.8
50–74	2	2.4	4	4.2	25	21.7	12	13.9	4	4.8
75–99	4	1.9	3	3.1	15	19.3	14	13.4	8	6.3
100+	2	2.2	3	4.4	14	13.4	17	14.5	11	12.5

Fig. 6.3 Log relative risk of oesophageal cancer according to five levels of alcohol consumption

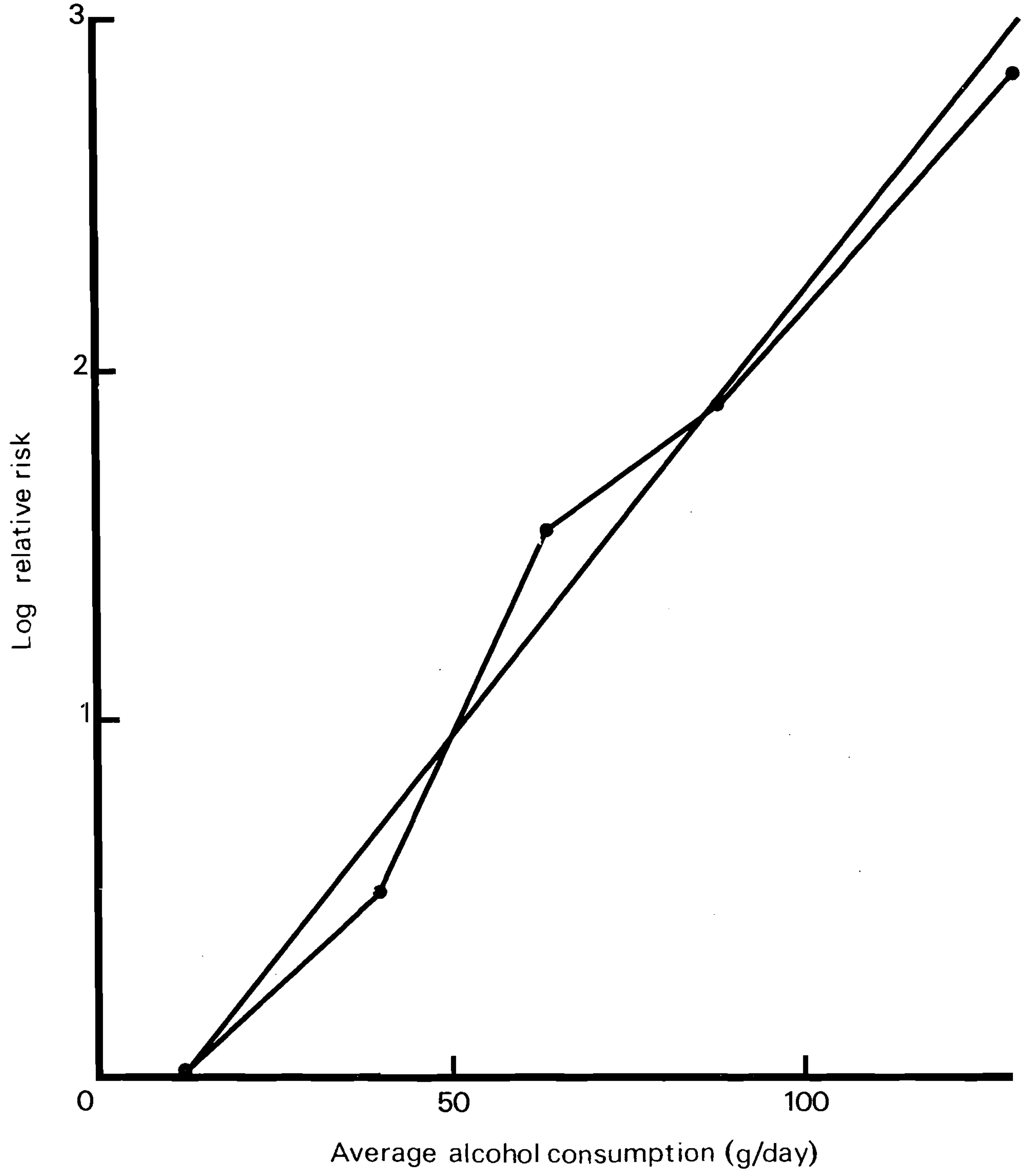
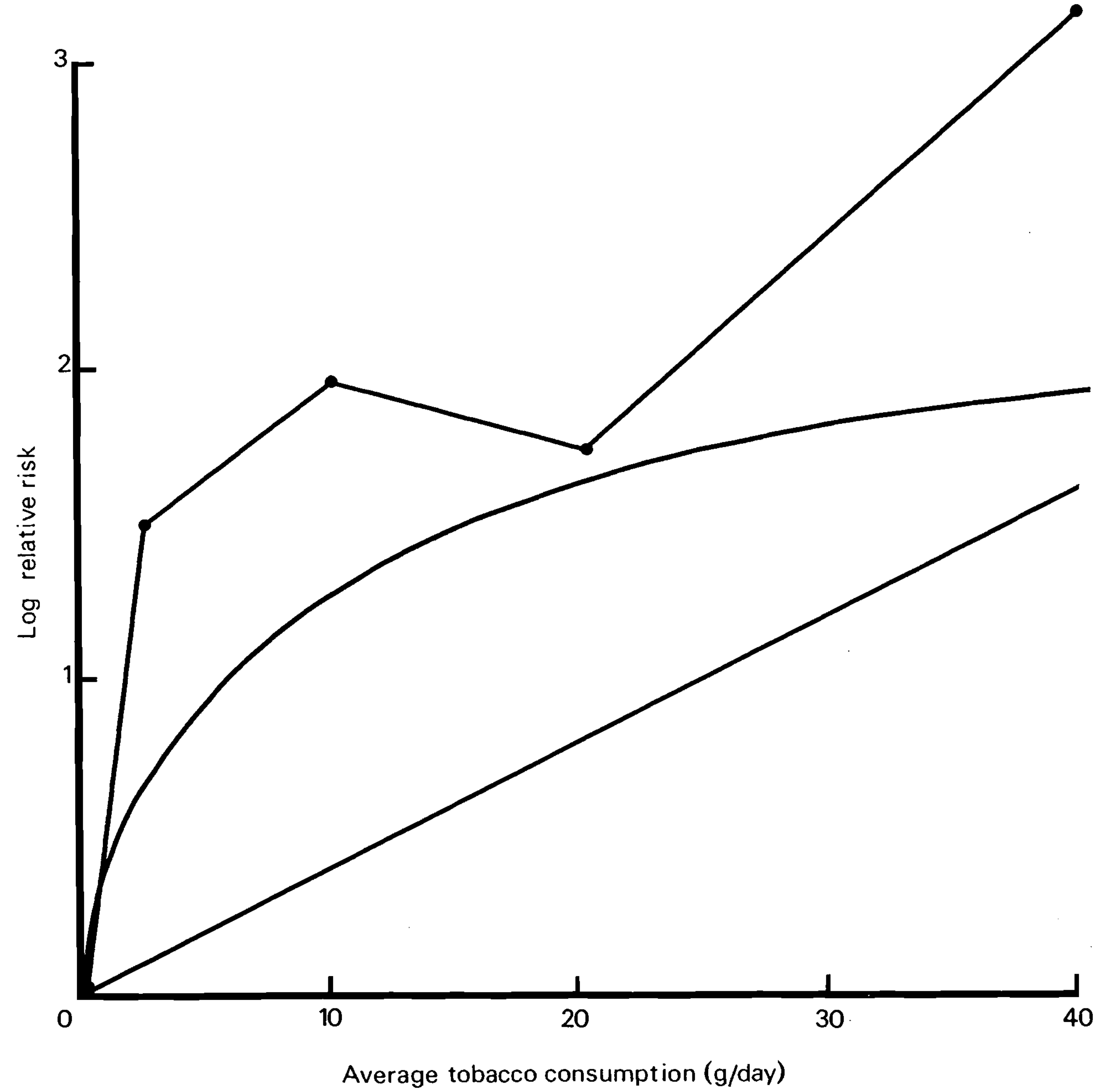


Fig. 6.4 Log relative risk of oesophageal cancer according to five levels of tobacco consumption



6.7 Quantitative analysis of grouped data

An important feature of the Ille-et-Vilaine data which was ignored in the preceding section is that different levels of the two risk factors have a prescribed order. It is possible to assign to each of them a quantitative value of exposure, for example the midpoint of the respective interval, or the average over the sample of the values of the underlying continuous variable within that interval. Natural values to assign to the four levels of alcohol are 20, 60, 100 and 150 g/day (Figure 6.1), which are interval midpoints except that 150 represents the approximate median of the values in the last open-ended interval above 120 g/day. Similarly, for tobacco, natural values are 5, 15, 25 and 40 g/day (Figure 6.2). Even the stratification variable is quantitative, with equally-spaced intervals centred about 30, 40, 50, 60, 70 and 80 years of age.

Quantitative aspects of the data may be accounted for in the analysis by using continuous regression variables in place of the categorical ones. There are several advantages in this approach. First, the data can often be adequately summarized by a smaller number of parameters, which facilitates interpretation. Tests for the significance of individual regression coefficients are single degree of freedom tests for trend which, as we have repeatedly emphasized, are generally more powerful than tests directed against global alternatives to the null hypothesis. This feature is especially important in exploring possible interactions, since chi-square statistics based on qualitative interaction variables tend to have rather large numbers of degrees of freedom. Quantitative interaction variables, obtained as the product of the quantitative variables representing the main effects of the corresponding factors, enable us to identify particular patterns of departure from the basic linear model.

Suppose for the moment that a single risk factor has been divided into K levels corresponding to values $x_1, \ldots, x_K$ of a quantitative variable. Cases and controls may be classified into one of IK cells on the basis of stratum (i) and risk factor (k). A partial selection of logistic regression models which would be appropriate to fit to the disease probabilities $P_i(x_k)$ may be outlined as follows:

Model equation logit $P_i(x_k)$ =	No. of independent parameters	Goodness-of-fit statistic G	Interpretation/Description
α_i	I	G_1	Relative risk of unity in all strata: no effect of risk factor
$\alpha_i+\beta_1 x_k$	I+1	G_2	Linear increase in log-relative risk with exposure, same slope for each stratum
$\alpha_i+\beta_1 x_k+\beta_2 x_k^2$	I+2	G_3	Quadratic effect of exposure on log-relative risk
$\alpha_i+\beta_i x_k$	2I	G_4	Linear effect of exposure, but slope varies depending on stratum
$\alpha_i+\beta_k$	I+K−1	G_5	Individual relative risk for each exposure level
$\alpha_i+\beta_k+\gamma_{ik}$	IK	$G_6 = 0$	No constraints at all: separate relative risks in each stratum

$\beta_k = \gamma_{ik} = 0$ by convention for k = 1 and all i.

By comparing the statistics corresponding to different models one may test several hypotheses. For example G_1–G_5, a χ^2_{K-1} statistic, provides an unstructured (qualitative) test for the effects of the risk factor like those considered in the last section. *Its value would not be changed by a re-ordering of the exposure categories.* G_1–G_2, which has but a single degree of freedom, yields a much more specific test for linear trend in log relative risk with increasing exposure. A global test of departures from the linear model is provided by G_2–G_5, on K–2 degrees of freedom, while G_2–G_3 is a χ_1^2 statistic specifically designed to test for curvature in the regression line. Finally, G_2-G_4 is a χ_{I-1}^2 statistic testing the parallelism of the regression lines in the I strata. Lack of parallelism means that the relative risks for different exposure levels vary from one stratum to another, i.e., that there are interactions between stratification variables and risk factors. Notice that the goodness-of-fit statistic for model 6 is 0. Since the number of independent parameters equals the number of observations, there is perfect agreement between model and data in this case.

A similar but somewhat more elaborate set of models was fitted to the 96 grouped data records from Ille-et-Vilaine, treating the two risk factors alternately as qualitative and quantitative variables at four levels each. The values assigned to each level are as indicated above, namely 20, 60, 100 and 150 g/day for alcohol and 5, 15, 25 and 40 g/day for tobacco. In fact these x values were not used in the regression analyses in their original form, since this would have led to computational problems, especially with the square terms. Instead, alcohol consumption was expressed in units of 100 g/day, with values 0.2, 0.6, 1.0 and 1.5, while tobacco was expressed in units of 10 g/day. It is sometimes helpful to go even further and to *standardize* all regression variables, i.e., scale and centre them so that they have approximate mean values of zero and variances of unity, before proceeding with the numerical analyses.

Table 6.10 summarizes the results. In identifying the various models we have used the following shorthand: ALCGRP and TOBGRP denote the qualitative effects of alcohol and tobacco, each representing three indicator regression variables; ALC and TOB are single variables which represent the quantitative effects. All models contain the six stratum parameters α_i which express the qualitative effects of age. Model 1 is identical with Model 4 of Table 6.5, both alcohol and tobacco consumption being treated as qualitative factors which combine multiplicatively.

Comparing Models 1 and 2 there is some slight evidence that the increase in log relative risk with alcohol may not be purely linear ($\chi_2^2 = 5.07$, $p = 0.08$); however, since the specific test for curvature obtained by comparing Models 2 and 3 is not at all significant ($\chi_1^2 = 0.11$, $p = 0.95$), we feel reasonably confident in attributing these deviations from a straight line relationship to chance. Linearity of the trend with tobacco, at least as based on the grouping into four levels, seems quite adequate; compare Model 4 with Models 1 and 5. Thus, Model 6, containing just one term for each of alcohol and tobacco, fits the data nearly as well as a model with four more parameters representing the non-linear effects of the two risk variables. From the regression coefficients[1] for Model 6, $\hat{\beta}_{ALC} = 2.55$ and $\hat{\beta}_{TOB} = 0.409$, we estimate that the risk of oesophageal cancer increases by a factor of exp(0.255) = 1.29 for every additional 10 grams of alcohol consumed per day, and by exp(0.409) = 1.51 for each additional 10 grams of tobacco.

Model 7 contains a quantitative term representing the linear × linear interaction of alcohol and tobacco. A significant value for its coefficient would have indicated a trend in the slope of the alcohol relationship with increasing consumption of tobacco, or

[1] Remember that for these calculations alcohol was expressed in units of 100 g/day and tobacco in 10-gram units.

Table 6.10 Results of fitting various logistic models with qualitative and quantitative regression variables to grouped data from the Ille-et-Vilaine study of oesophageal cancer

Model	Parameters fitted (in addition to stratum or age effects)	DF	Goodness of fit G	Hypothesis tested/interpretation ,
1	ALCGRP + TOBGRP	76	82.34	Multiplicative model with qualitative risk variables
2	TOBGRP + ALC	78	87.41	Linear effect of alcohol
3	TOBGRP + ALC + ALC^2	77	87.01	Linear + quadratic effects of alcohol
4	ALCGRP + TOB	78	84.53	Linear effects of tobacco
5	ALCGRP + TOB+ TOB^2	77	83.73	Linear + quadratic effects of tobacco
6	ALC + TOB	80	89.02	Linear effects of alcohol and tobacco
7	ALC + TOB + ALC × TOB	79	88.05	Linear × linear alcohol/tobacco interaction
8	ALCGRP + TOBGRP + ALC × TOB	75	81.37	Linear × linear alcohol/tobacco interaction in qualitative model
9	ALCGRP + TOBGRP + ALC × AGE	75	80.08	Linear increase in slope of alcohol effect with age
10	ALCGRP + TOBGRP + TOB × AGE	75	82.33	Linear increase in slope of tobacco effect with age

KEY:

ALCGRP	= indicator variables for alcohol levels
TOBGRP	= indicator variables for tobacco levels
AGE	= quantitative age variable
ALC	= quantitative alcohol variable
TOB	= quantitative tobacco variable

equivalently a trend in the tobacco relationship with alcohol. However, there is no evidence for such a trend ($\chi_1^2 = 0.97$, $p = 0.32$). Model 8 illustrates that quantitative interaction terms may be used even when the model expresses the main effects qualitatively. Subtracting G_8 from G_1 leads to a nearly identical test for the quantitative alcohol × tobacco interaction ($\chi_1^2 = 0.97$, $p = 0.32$). Quantitative interaction variables may be quite valuable in giving some specificity to the search for interactions even if one does not want to assume a particular form for the main effects.

The last two models search for similar quantitative interactions with the stratification variable. A negative regression coefficient for the ALC × AGE term in Model 9 indicates a tendency for the alcohol relative risk to diminish with advancing age. However, it is not a significant trend ($\chi_1^2 = 2.28$, $p = 0.13$). There is no indication at all of a systematic change in the tobacco effect with age. Thus our previous conclusions based on the qualitative analysis of interactions are in this example further supported by the quantitative approach.

6.8 Regression adjustment for confounders

Stratification, whether in the context of M-H methodology or logistic regression, has traditionally been used to control the confounding effects of nuisance factors. Typically, we define a separate stratum for each combination of levels of the nuisance factors, assigning to each one a parameter in the model. If there are several such factors, or if they occur at very many levels, the total number of strata can become quite large. For example, with three stratifying factors at 3, 4 and 5 levels, respectively, the total number of α parameters in (6.2) is $3 \times 4 \times 5 = 60$. Since the available data usually place severe limitations on the number of strata which may be incorporated in the analysis, alternative methods for the control of confounding must be considered.

From the discussion in § 6.1 it is clear that the *practice of stratification is tantamount to saturating the effects of the nuisance factors with parameters.* Not only the main effects, but also all the first and higher order interaction terms are represented. This practice is unnecessary, however, unless we have good reason to believe that such higher order interactions are present. An obvious alternative to stratification for the control of confounding variables is to incorporate their effects directly into the model. This allows us much more flexibility in deciding which of the higher order interaction terms to retain and which to discard. The approach may be especially efficacious with continuous nuisance factors whose effects can be adequately summarized in a few quantitative regression variables.

This does not mean, however, that risk and nuisance variables are treated symmetrically in the analysis. For risk factors our goal is to identify the most important ones and quantify their influence in a precise and meaningful way. This implies that we *economize* on the number of parameters used to represent them and that we retain in the multivariate risk equation only those which have reasonably significant effects.

For nuisance factors, on the other hand, the effects on disease have presumably already been conceded, or in any event are not the specific concern of the study. They are included only to ensure that the estimates of relative risk are free from possible confounding effects, and no specific meaning is to be attached to their coefficients. *Hence, known confounding variables should be included in the equation regardless of statistical significance if such inclusion changes the estimated coefficients of the risk variables by any appreciable degree* (§ 3.4).

We illustrate the regression adjustment for confounding effects with the grouped data from Ille-et-Vilaine, specifically the age adjustment of estimates in the qualitative multiplicative model (Model 4, Table 6.5). Table 6.11 compares the previous estimates, obtained using stratification in six age groups, to estimates for which quantitative adjustments were made by introducing polynomial expressions in age group into the equation. Let i denote the age stratum, j the alcohol group, and k the tobacco group. The left-hand column presents the unadjusted estimates, based on an equation of the form

$$\text{logit } P_{ijk} = \alpha_0 + \beta_j(\text{alc}) + \beta_k(\text{tob})$$

where age does not appear at all. The next column shows the changes in the alcohol and tobacco coefficients upon introduction of a single linear term in age

$$\text{logit } P_{ijk} = \alpha_0 + \alpha_1 i + \beta_j(\text{alc}) + \beta_k(\text{tob}).$$

The third column shows the effect of adding a quadratic age term $\alpha_2 i^2$, and so on.

Comparing G's for the extreme left- and right-hand columns of Table 6.11 it is clear that age has an enormous influence on risk ($\chi_5^2 = 126.5$). Nevertheless, the differences between these two columns in the relative risk estimates for alcohol and tobacco are rather minor, which implies that the confounding effects of age are quite weak. The explanation for this phenomenon has been given in § 3.3. While age is strongly related to risk, it has only a weak correlation with the level of exposure to alcohol and tobacco (Table 4.2) and hence would not be expected to be a strong confounder.

Inclusion of a single linear term in age group results in an enormous improvement in overall fit and brings the estimated coefficients quite close to those obtained *via* stratification. Fitting both linear and quadratic terms yields results which are virtually identical to those obtained with higher degrees of adjustment. These comparisons, which are typical of our experience with quantitative nuisance factors, indicate that effective control of confounding is often obtainable by inclusion of a few polynomial terms in the regression equation, thus obviating the need for stratification. The regression method of adjustment should generally work well unless disease incidence or the exposure to other risk factors depends in a complicated, non-linear way on the nuisance variables.

Table 6.11 Estimate of log relative risk for each alcohol and tobacco category according to the degree of adjustment for age: Ille-et-Vilaine oesophageal cancer study

Risk category	Type of analysis					
	Unadjusted	Polynomial adjustment for age group				Stratification in
		Linear	Quadratic	Cubic	Quartic	six age groups
Tobacco (g/day)						
0–9	0.0	0.0	0.0	0.0	0.0	0.0
10–19	0.39	0.46	0.44	0.43	0.43	0.44
20–29	0.43	0.55	0.51	0.50	0.50	0.51
30+	0.99	1.52	1.63	1.63	1.64	1.64
Alcohol (g/day)						
0–39	0.0	0.0	0.0	0.0	0.0	0.0
40–79	1.23	1.53	1.44	1.44	1.44	1.44
80–119	2.00	2.17	1.99	2.00	1.99	1.98
120+	3.18	3.60	3.57	3.58	3.59	3.60
Goodness-of-fit statistic G	208.8	101.9	84.6	84.0	83.8	82.3
Degrees of freedom	81	80	79	78	77	76

6.9 Analysis of continuous data

The full power of the regression approach to case-control studies is obtained when continuous risk variables are analysed in the original form in which they were recorded, rather than by grouping into intervals whose endpoints are often arbitrarily chosen. This permits the incorporation of many more variables than would be possible using grouped data, their joint effects being summarized by a relatively small number of parameters. Of course such an increase in power and flexibility is not without associated costs. Perhaps the most serious are potential errors in the estimated relative risks arising from a *mis-specification* of the model. Careful exploration of the adequacy of the postulated relationships is essential to avoid over-interpretation of the data. Transformations and interaction terms should be used where required to improve the fit.

Another cost associated with the use of continuous risk variables is monetary. Since individual data records for each subject must be processed repeatedly during the iterative fitting process, large amounts of computer time can be required to analyse a comprehensive series of models. With the Ille-et-Vilaine study, for example, only 88 data records were required for the grouped data analyses of § 6.6 and 6.7. All 975 records, one for each subject, were needed for the continuous analysis, and computer costs for fitting equivalent models were 5–10 times higher. Of course additional information is contained in the original, continuous data which is undoubtedly worth the price of extraction, especially when one considers that costs of data processing and analysis are only a small part of the total cost of any study.

In the first series of continuous models fitted to the Ille-et-Vilaine data we used quantitative variables representing alcohol and tobacco consumption as well as various transformations of these. "Alcohol" (ALC) was a true continuous variable in that it took on 163 separate values between 0 and 268 g/day (inclusive) among the 975 study subjects. 'Tobacco" (TOB), on the other hand, had been recorded as a discrete variable with nine levels. For the analyses reported here quantitative values were assigned to each such level, as they had been earlier for the grouped data analyses:

Coding of quantitative tobacco variable

Level	Interval (g/day)	Assigned value (x)
1	0	0.0
2	1–4	2.5
3	5–9	7.5
4	10–14	12.5
5	15–19	17.5
6	20–29	25.0
7	30–39	35.0
8	40–49	45.0
9	50+	60.0

As an alternative to using ALC and TOB as linear terms in the model, transformations of each of these were considered. A particularly appropriate transformation for variables which represent dose rates of continuous exposures is the log transform. Postulating a log-linear relation of the form log RR(x) = $\alpha + \beta$ log(x) means that risk itself is proportional to a power of dose, x^β, a relationship known to occur frequently from both human and animal studies (see § 6.11). Since both ALC and TOB took on 0 values it was necessary to "start" the logs by adding 1 to each before transforming it, in order to avoid infinities. Note that with either the original (ALC and TOB) or the transformed [LOG(ALC + 1) and LOG(TOB + 1)] variables, non-consumers of both tobacco and alcohol are automatically assigned relative risks of 1.0. This is because the values of all risk variables are 0 for individuals consuming no alcohol and no tobacco.

Table 6.12 presents the results. The first model, which includes linear terms for each of alcohol and tobacco, may be compared with Model 6, Table 6.10, of the grouped data analysis. Agreement between the two sets of coefficients is remarkably good: the log relative risk is estimated to increase by 0.255 (grouped) or 0.260 (continuous) for every additional 10 grams of alcohol, while for 10 grams of tobacco the corresponding figures are 0.409 and 0.405.

In contrast to the situation with grouped data, the goodness-of-fit statistics shown in the fourth column of Table 6.12 should not be interpreted as chi-squares with the indicated degrees of freedom. Because the number of cases in each "group" is 0 or 1 according to whether the record refers to a case or control, a direct comparison of observed and expected numbers is not helpful in determining the adequacy of the model. Instead the *differences* between the measures for nested models evaluate their relative goodness of fit, as explained in § 6.4.

Table 6.12 Logistic regression analysis of continuous risk variables: Ille-et-Vilaine oesophageal cancer study

Model	No. of parameters[a]	DF	Goodness of fit G	Regression coefficients for each risk variable (standardized coefficients in parentheses)[b]						
				ALC	TOB	LOG (ALC+1)	LOG (TOB+1)	ALC^2	TOB^2	LOG^2 (TOB+1)
1	8	967	695.4	0.0260	0.0405					
				(10.00)	(5.13)					
2	8	967	749.5		0.0411	0.933				
					(5.47)	(7.02)				
3	8	967	683.2	0.0252			0.539			
				(9.66)			(9.33)			
4	8	967	734.8			0.890	0.555			
						(6.73)	(6.19)			
5	9	966	693.9	0.0257	0.0648				−0.0006	
				(9.88)	(3.05)				(1.24)	
6	9	966	682.8	0.0202			0.539	0.0033		
				(2.52)			(5.77)	(0.65)		
7	9	966	681.1	0.0251			0.965			−0.114
				(9.61)			(3.05)			(1.44)

[a] Includes the six age terms α_i in addition to the alcohol and tobacco parameters shown

[b] Both ALC and TOB are expressed in units of g/day.

For Model 2 of Table 6.12, the effect of alcohol consumption is expressed on a logarithmic rather than an arithmetic scale. In view of the marked decrease in the log likelihood, the log scale is clearly not appropriate for alcohol. On the other hand, the fit is substantially improved when the effect of tobacco is expressed in this way (Model 3). Addition of square terms in ALC (Model 6) or LOG(TOB+1) (Model 7) do not result in a statistically significant improvement over the model containing these two variables alone ($G_3 - G_6 = 0.4$, $p = 0.5$; $G_2 - G_7 = 2.1$, $p = 0.15$). It is of interest to note that not even use of both linear and quadratic terms in TOB (Model 5) achieves the goodness of fit produced by expressing this variable on a log scale.

Taking ALC and LOG(TOB+1) as the basic risk variables, tests were made for interaction effects between these two factors, as well as between each of them and age. Addition of an ALC×LOG(TOB+1) interaction term to the model reduced the goodness-of-fit statistic very little, to 682.6 ($\chi_1^2 = 0.6$, $p = 0.4$). Likewise, no interactions of alcohol with age ($\chi_1^2 = 1.1$, $p = 0.3$) nor of tobacco with age ($\chi_1^2 = 0.2$, $p = 0.7$) were apparent. Thus the quantitative regression analysis of the continuous data confirms the lack of interaction effects noted previously in our analysis of the grouped data.

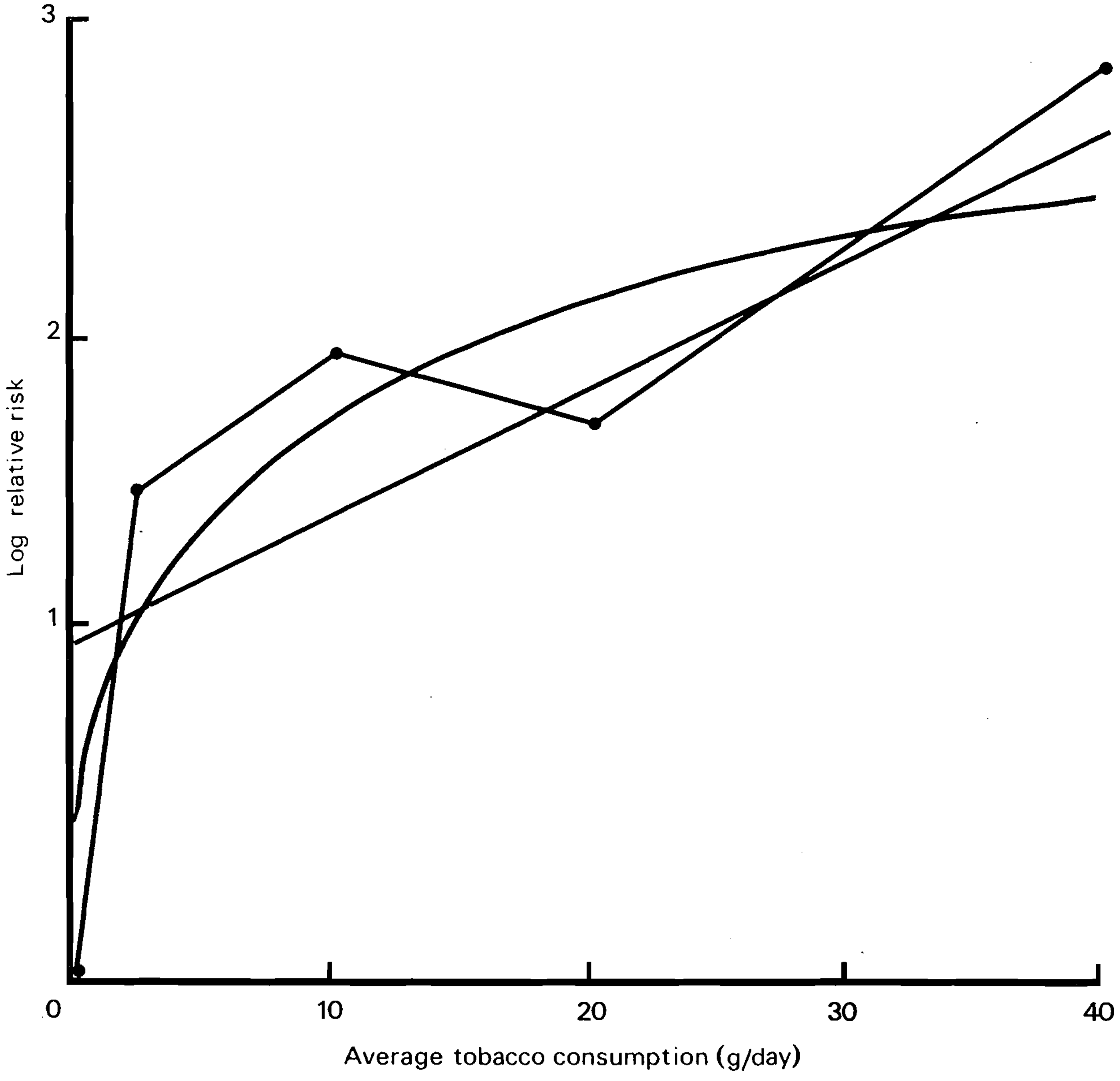
In summary, the changes in risk of oesophageal cancer associated with increased alcohol and tobacco consumption are well represented by a model in which the effects of the two factors combine multiplicatively. The proportional increase in risk accompanying additional quantities of alcohol and tobacco, expressed in units of g/day, is estimated to be

$$(\mathrm{TOB}+1)^{0.54}\exp(0.025\times \mathrm{ALC}).$$

Standard errors of the regression coefficients, 0.0026 for alcohol and 0.058 for tobacco, may be used to put approximate confidence limits about the estimates. Dividing the standard errors into the coefficients themselves yields the standardized values (Table 6.12), which may be referred to tables of the normal distribution to test for the significance of individual terms in the regression equation. Clearly both alcohol and tobacco have highly significant independent effects, as has already been established using other methods.

A plot of the estimated linear increase in log relative risk with alcohol (Figure 6.3) shows excellent agreement with the results of the qualitative analyses. Similar plots for tobacco are shown in Figure 6.4. Here the situation at first appears somewhat paradoxical. The estimated relative risks from the qualitative analysis lie entirely above those based on the log transform, which in turn lie above those derived from the linear model. The explanation for this apparently bizarre phenomenon is not hard to find. It is due to the arbitrary selection of 0 as a baseline value for tobacco, which constrains all three curves to pass through the origin of the graph. Any other value for tobacco could just as well have been chosen as baseline and assigned a 0 log relative risk, in which case the curves would all be displaced so as to pass through 0 at that point. In other words *the origin of the scale of log-relative risk is completely arbitrary and it is only the shapes of the curves which have any meaning.* To compare and contrast these shapes better, Figure 6.5 shows the same three curves except that the linear curve has been displaced upwards 0.96 units and the log curve up 0.48 units. The superior fit of the model using the log term is evident from this graph.

Fig. 6.5 Log relative risk of oesophageal cancer according to five levels of tobacco consumption: not constrained to pass through origin



We conclude this section with an illustration of the ability of the logistic regression model to investigate the simultaneous effects of a large number of continuous risk variables. In order to estimate the average daily amount of alcohol consumed by each study subject, interviewers posed separate questions regarding the pattern and frequency of use of wine, beer, cider, aperitifs and digestives. The last two categories included distilled beverage such as whisky (an aperitif) and brandy (a digestive). Separate variables representing the average daily consumption of alcohol in each form were available in the computer file. These had been obtained from the reported amounts drunk by consideration of the usual alcoholic content: 8% by weight for wine; 3% for beer and so on. Table 6.13 shows the distribution of each of the five beverage variables separately for cases and controls. Note that the sums of their mean values equal the means for alcohol (Table 4.1), as they should since ALC is obtained as the total of the component variables. The contributions from wine and cider are of roughly equal importance and those for beer and digestives, while lesser, are certainly not negligible. However, since so few people in this population report that they consume more than a few grams per day of aperitifs, we are already aware that it may be impossible to evaluate aperitifs as a separate risk factor.

Correlations among the five beverage variables, and of these with age and tobacco, are presented in Table 6.14 for the control population. The lack of strong correlations with age and tobacco inform us that these two variables are unlikely to confound the beverage effects to any appreciable degree. Even among the beverage variables the correlations are relatively weak, the strongest being between cider and digestives (ϱ = 0.31). Evidently cider drinkers tend to consume less wine, beer and aperitifs, but more digestives, than non-cider drinkers.

The rationale for using the summary alcohol variable in the statistical analysis, as done earlier, is the belief that the alcohol content of the beverages is responsible for the apparent association with oesophageal cancer and not some other characteristic such as impurities. In order to evaluate this hypothesis we fitted a series of models in which five separate beverage variables were used in place of total alcohol. The results

Table 6.13 Distribution of average daily amounts of alcohol consumption by type of beverage, for cases and controls: Ille-et-Vilaine study of oesophageal cancer

Average daily amount (g)	Type of beverage									
	Beer		Cider		Wine		Aperitif		Digestive	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
0	141	493	76	373	28	142	105	346	76	430
1–9	21	144	12	64	30	214	90	419	55	236
10–39	18	103	49	229	63	293	5	10	58	98
40–79	11	30	39	92	52	104	0	0	10	10
80+	9	5	24	17	27	22	0	0	1	1
Mean	9.1	5.7	30.7	15.5	34.3	17.8	1.1	1.1	9.7	4.3
SD	22.7	13.7	37.4	21.6	37.1	21.2	2.6	2.0	15.1	8.9

Table 6.14 Correlations between alcoholic beverage variables, tobacco and age in control population: Ille-et-Vilaine study of oesophageal cancer

	Age	Tobacco	Beer	Cider	Wine	Aperitif	Digestive
Beer	–0.18	0.20	1.00				
Cider	0.08	–0.10	–0.16	1.00			
Wine	–0.04	0.16	0.07	–0.27	1.00		
Aperitif	–0.09	0.15	0.09	–0.11	0.21	1.00	
Digestive	0.13	0.04	–0.03	0.31	–0.02	0.06	1.00

Table 6.15 Logistic regression analysis of continuous beverage variables: Ille-et-Vilaine oesophageal cancer study

Model	No. of parameters[a]	DF	Goodness of fit G	Regression coefficients for each risk variable (standardized coefficients in parentheses)						
				LOG (TOB+1)	ALC	Beer	Cider	Wine	Aperitif	Digestive
1	8	967	683.2	0.539 (9.33)	0.0252 (9.66)					
2	12	963	674.3	0.546 (5.70)		0.0252 (4.55)	0.0281 (7.33)	0.0312 (7.80)	–0.0660 (–1.48)	0.0120 (1.30)
3	11	964	676.6	0.535 (5.61)		0.0247 (4.48)	0.0284 (7.43)	0.0301 (7.70)		0.0109 (1.27)
4	10	965	678.2	0.537 (5.64)		0.0248 (4.52)	0.0299 (8.19)	0.0304 (7.83)		
5	8	967	793.2	0.592 (6.90)		0.0156 (3.09)				
6	8	967	766.4	0.690 (7.46)			0.0178 (5.86)			
7	8	967	766.3	0.536 (6.24)				0.0194 (5.89)		
8	8	967	802.6	0.627 (7.26)					–0.0109 (0.28)	
9	8	967	786.6	0.624 (7.16)						0.0291 (3.87)

[a] Includes six age parameters in addition to those shown

are shown in Table 6.15, of which the first line is simply a repeat of Model 3, Table 6.12 Model 2 shows that beer, cider and wine each have highly significant independent effects on the risk of oesophageal cancer. It is remarkable how close all three coefficients are to the 0.0252 estimated for total alcohol, which lends support to the idea that alcohol *per se* is responsible for the effect. On the other hand, the coefficients for the two distilled beverage categories are not significantly different from zero, and that for aperitifs is even negative.

Before jumping to the conclusion that the aperitifs and digestives have a lesser effect, or even no effect in proportion to their alcohol content, we should consider the data presented in Table 6.13. Since fewer people in the population consume large amounts of aperitifs or digestives there is less information available for evaluating their role, a fact which is reflected in higher standard errors for their coefficients in comparison with the other variables. The upper 95% confidence intervals for the log relative risks are 0.0216 for aperitifs and 0.0288 for digestives, and the latter at least is quite consistent with the range of values for beer, cider and wine. To test formally the hypothesis that the coefficients for all five beverage variables are equal we have merely to compare the goodness-of-fit statistics for Models 1 and 2. Since Model 1 uses the sum of the beverage variables as a single regression variable (ALC), it constrains the coefficients to be equal and is consequently contained in Model 2. The value of the test statistic is $G_1 - G_2 = 8.9$ which, when referred to tables of χ^2_4, gives $p = 0.06$, a result bordering on statistical significance.

To go one step further we can *partition* the χ_4^2 value into single degree of freedom components by considering two intermediate models: 1A, in which the coefficients for beer, cider and wine only were assumed equal; and 1B, in which all coefficients were assumed equal except for aperitifs. These yield goodness-of-fit statistics of $G_{1A} = 678.9$ and $G_{1B} = 675.3$. Hence we may write $G_1 - G_2 = (G_1 - G_{1A}) + (G_{1A} - G_{1B}) + (G_{1B} - G_2)$, i.e., $8.9 = 4.3 + 3.6 + 1.0$, partitioning the χ_4^2 statistic into two χ_1^2's and one χ_2^2. The first, $G_1 - G_{1A}$, tests whether aperitifs have an effect different from the average of the remaining beverages ($p = 0.04$); the second, $G_{1A} - G_{1B}$, whether digestives differ from the remaining three ($p = 0.06$); and the last, $G_{1B} - G_2$, tests for differences among the coefficients of beer, cider and wine ($p = 0.60$). But, since the particular partitioning was suggested by the data rather than from *a priori* considerations, we are faced with *a multiple comparisons* dilemma and should discount the observed p values.

In the last analysis the situation is somewhat ambiguous. While digestives appear to have lesser effects than the other variables, and aperitifs no effect at all, we cannot rule out at conventional levels of statistical significance the possibility that all beverages contribute to the risk in proportion to their alcohol content.

6.10 Interpretation of regression coefficients

The preceding example considered a model with 12 independent parameters, each of which had a reasonably clear and straightforward interpretation. Six of the parameters, the α's attached to the six age strata, were included only to account for possible confounding effects of age. Since age effects were not of special interest, their estimates were not even presented in Table 6.15. However, the controls were obtained as a reasonably random sample of the adult male population so that differences between the α's could be interpreted in terms of log relative risks for the corresponding age groups. (From the α coefficients in Table 6.2, for example, it appears that risk does not change much with age beyond 55 years.) On the other hand, had the sample been stratified by design on the basis of age, no meaning at all could be attached to the α parameters since the effects of age on risk would then be completely confounded with the sampling fractions for different ages (§ 6.3).

While there is generally little interest in the actual values taken on by the α estimates, apart from knowing that the variables they represent have been "accounted for", this is hardly true for the β's. These we have repeatedly interpreted as indicating the change in risk associated with changes in the corresponding regression variables. It is a little disconcerting, therefore, to realize that the *estimated regression coefficients may change drastically according to what other variables are included in the model.* Such changes are to be anticipated whenever there is *collinearity* among the regression variables, meaning simply that their values tend to be correlated in the sampled data. Mosteller and Tukey (1977) provide a good discussion of this problem, which is fundamental to all regression models. Here we consider a few of the main issues, mostly by means of example.

In the Ille-et-Vilaine data there was a remarkable lack of collinearity among age and the levels of consumption of tobacco and total alcohol (Table 4.2). Consequently the estimated relative risks associated with each of these factors were little affected by which others were accounted for in the equation. For example, the (age-adjusted) relative risks for the four alcohol categories were 1.0, 4.2, 7.4 and 39.7 without inclusion of tobacco in the analysis, and 1.0, 4.2, 7.2 and 36.6 with such inclusion (Table 6.6).

A better illustration of the effects of collinearity is provided by the analyses of the contributions of individual alcoholic beverages (Table 6.15). We note first that neither the coefficients nor the standard errors of beer, cider and wine are much affected by the presence or absence of aperitif or digestive in the equation (Models 2–4), provided all three of the alcoholic beverages with significant effects are included. One would anticipate such a result if either (1) there was no correlation between the beverage variables, or (2) aperitif and digestive had no effect on risk beyond that explained by such a correlation (§ 3.4). However, when any one of beer, cider or wine is used as the *only* alcohol variable (Models 5–7), its coefficient and degree of statistical significance are noticeably reduced. This reflects the fact that cider is *negatively* correlated with both wine ($\varrho = -0.27$) and beer ($\varrho = -0.16$). Since an individual consuming a large amount of cider tends to consume less than the average amount of the other beverages, his apparent cancer risk relative to someone who drinks no cider is reduced unless the effects of these other beverage variables are accounted for by inclusion in the equation.

A different type of change occurs when digestive is used as the only alcohol variable (Model 9). Here the coefficient increases markedly from its value when all alcoholic beverage variables are included, and attains an apparently high level of statistical significance. The explanation now is the *positive* correlation of digestive with cider ($\varrho = 0.31$), such that when cider is not included in the equation, digestive serves, at least partially, as a *proxy* for its effects. After accounting for the effects of cider the coefficient for digestive falls to a non-significant level. On the other hand, since the correlations of digestive with beer and wine are essentially zero (Table 6.14), one would not expect the digestive coefficient to be much altered by the presence of these latter two variables.

Collinearity is bound to arise when both a variable and its square are included in the same equation. Compare, for example, Model 3 with Model 7 in Table 6.12. Introduction of the square term in LOG(TOB + 1) results in an almost doubling of the coefficient for the linear term, from 0.539 to 0.965. At the same time the standardized value decreases, from 9.33 to 3.05, indicating a roughly sixfold increase in the standard error. This is true in spite of the fact that the coefficient of the added variable, LOG^2(TOB+1), is not statistically significant at all. Indeed, if we were to evaluate the significance of the tobacco effect only on the basis of the standardized coefficients in the quadratic model, we would be sorely misled. The significance of the trend in risk with increased tobacco consumption is well expressed by the single linear term in Model 3; and the large standard errors for LOG(TOB + 1) and LOG^2(TOB + 1) in Model 7 tell us not that these variables are unimportant, but rather that there are many different sets of coefficients for them which express more or less equally well the relationship found in the data. This example illustrates that there is little point in trying to interpret individual coefficients and standard errors in a polynomial regression. A plot of the fitted relationship over the range of the regression variables conveys a much more accurate impression of what the equation means.

A similar type of artificial association can arise between one variable representing the main effects of a factor and others representing its interactions, at least if care is not taken in how these interactions are coded. For example, in order to investigate the interaction between age and alcohol we added to Model 3 of Table 6.12 a variable ALC × AGEGRP, where AGEGRP took on the values 1 to 6 of the age group. Although this improved the goodness of fit only slightly, from $G = 683.2$ to $G = 682.1$, and the interaction terms had a non-significant regression coefficient, its inclusion in the equation markedly affected the coefficient of ALC. The estimated regression equation (ignoring age effects) changed from

$$\underset{(9.66)}{0.0252\,\mathrm{ALC}} + \underset{(9.33)}{0.539\,\mathrm{LOG(TOB{+}1)}}$$

to

$$\underset{(3.58)}{0.0348\mathrm{ALC}} + \underset{(5.76)}{0.536\mathrm{LOG(TOB{+}1)}} - \underset{(-1.04)}{0.00246\mathrm{ALC}\times\mathrm{AGEGRP}},$$

and by comparing the standardized coefficients (shown in parentheses), we see that the standard error of ALC increased from 0.00261 to 0.00972. Again the explanation is the high degree of collinearity between ALC and ALC × AGEGRP, which can be substantially reduced by subtracting from AGEGRP its *modal* value of 4 before multiplying. This leads to an equation

$$\underset{(9.56)}{0.0250\mathrm{ALC}} + \underset{(5.76)}{0.536\mathrm{LOG(TOB{+}1)}} - \underset{(-1.04)}{0.00246\mathrm{ALC}\times(\mathrm{AGEGRP\text{-}4})}$$

which represents *exactly the same relationship* as the previous one. However, because the main effect and interaction variables have been coded to reduce the correlation between them, the changes in the coefficient and standard error of the main effect variable are much reduced. Routine coding of interaction or cross-product variables by subtracting mean or modal values from their component parts before multiplying is recommended to avoid the anomalies provoked by such artificial collinearity.

A less artificial example of high correlation between two regression variables occurs when both are measuring the same fundamental quantity in a somewhat imperfect way. In attempting to relate arsenic exposure to cancer risk, for example, we might determine the arsenic concentration of both fingernails and hair of cases and controls, and use each as an indicator of chronic exposure. If these two measures turned out to be highly correlated, as they would if both were good indicators of long-term exposure, it would make little sense to attempt to evaluate their separate effects on risk by including them both in the regression equation. Instead we would take an average or composite of the two values as a single measure of arsenic exposure, and use this along with variables representing other risk factors.

Of course in some problems the collinearity between regression variables will reflect a real association between the corresponding risk factors in the population. While regression analysis is the most powerful tool available for separating out the independent associations with risk, unambiguous answers are simply not possible when collinearity is high. In some cases a judgement as to which is the proper variable, or which risk factor is more likely to play a causal role, will dictate which variables to leave in the equation. If such a judgement cannot be made, one must simply admit that precise identification of the factor responsible for the effect is impossible. To quote from Mosteller and Tukey (1977): "We must be prepared for one variable to serve as a proxy for another and worry about the possible consequences, in particular, whether the proxy's coefficient siphons off some of the coefficient we would like to have on the proper variable, or whether a variable serves well only because it is a proxy. In either case, interpretation of the regression coefficient requires very considerable care." Much of the discussion in § 3.4 on whether or not one should adjust for apparent confounding variables is relevant here.

6.11 Transforming continuous risk variables

One of the more perplexing issues facing the analyst who uses quantitative regression methods is the choice of appropriate scales on which to express continuous risk variables. He must decide between original measurements, as recorded by machine or interviewer, and such transforms as logs, square roots, reciprocals, or any number of other possibilities. Since the object is to achieve a near-linear relationship between the quantitative regression variable and log risk, it usually helps to make some plots of relative risks for grouped data as we did in Figures 6.1 to 6.4. If the data are sufficiently extensive, so that a regular pattern emerges, one can at least rule out some of the possible choices on the grounds of lack of fit. For example it was fairly clear from both graphical and quantitative analysis of the Ille-et-Vilaine data that the effects of alcohol were best expressed on the original linear scale, while for tobacco a log transform was required.

However epidemiological data are rarely sufficient to enable fine distinctions to be made between rather similar functional forms for the dose-response relationship on statistical grounds alone. Accurate measurements of human exposure to potential risk factors are not often available. Hence recourse is made to animal experimentation for elucidating fundamental aspects of the carcinogenic process. Such experiments allow one to control fairly strictly the amounts of carcinogen administered to homogenous subgroups of animals, and data derived from them are more amenable to precise quantitative analysis than are data from observational studies of human populations.

Example: One animal model which has been used to suggest relationships for human epithelial tumours is that of skin-painting experiments in mice. In an experiment reported by Lee and O'Neill (1971), mice were randomly assigned to four dosage groups each containing 300 animals. Starting at about three weeks of age, benzo[*a*]pyrene (BP) was painted on their shaved backs in the following dosages:

Group 1	6 μg BP/week
Group 2	12 μg BP/week
Group 3	24 μg BP/week
Group 4	48 μg BP/week

The animals were examined regularly, and the week of tumour occurrence was taken to be the first week that a skin tumour was observed. Age-specific incidence rates of skin tumours were estimated for each dosage group according to the methods of § 2.1. The number of animals developing a skin tumour for the first time during any one week was divided by the number still alive and free of skin tumours at the middle of that week. Since few new tumours would arise in any given week, these age-specific estimates tended to be highly unstable. Consequently, the four dosage groups were compared in terms of the age-specific *cumulative* incidence rates (§ 2.3).

Figure 6.6 shows log-log plots of cumulative incidence against week. These are well described by four parallel straight lines, with distances between the lines for successive dosages roughly equal. In fact the cumulative incidence $\Lambda(t;\, x)$ of skin tumours which occur by week t, among animals receiving BP at dose x, is well described by the equation

$$\log \Lambda(t;x) = -17.6 + 1.78 \log(x) + 2.95 \log(t-18),$$

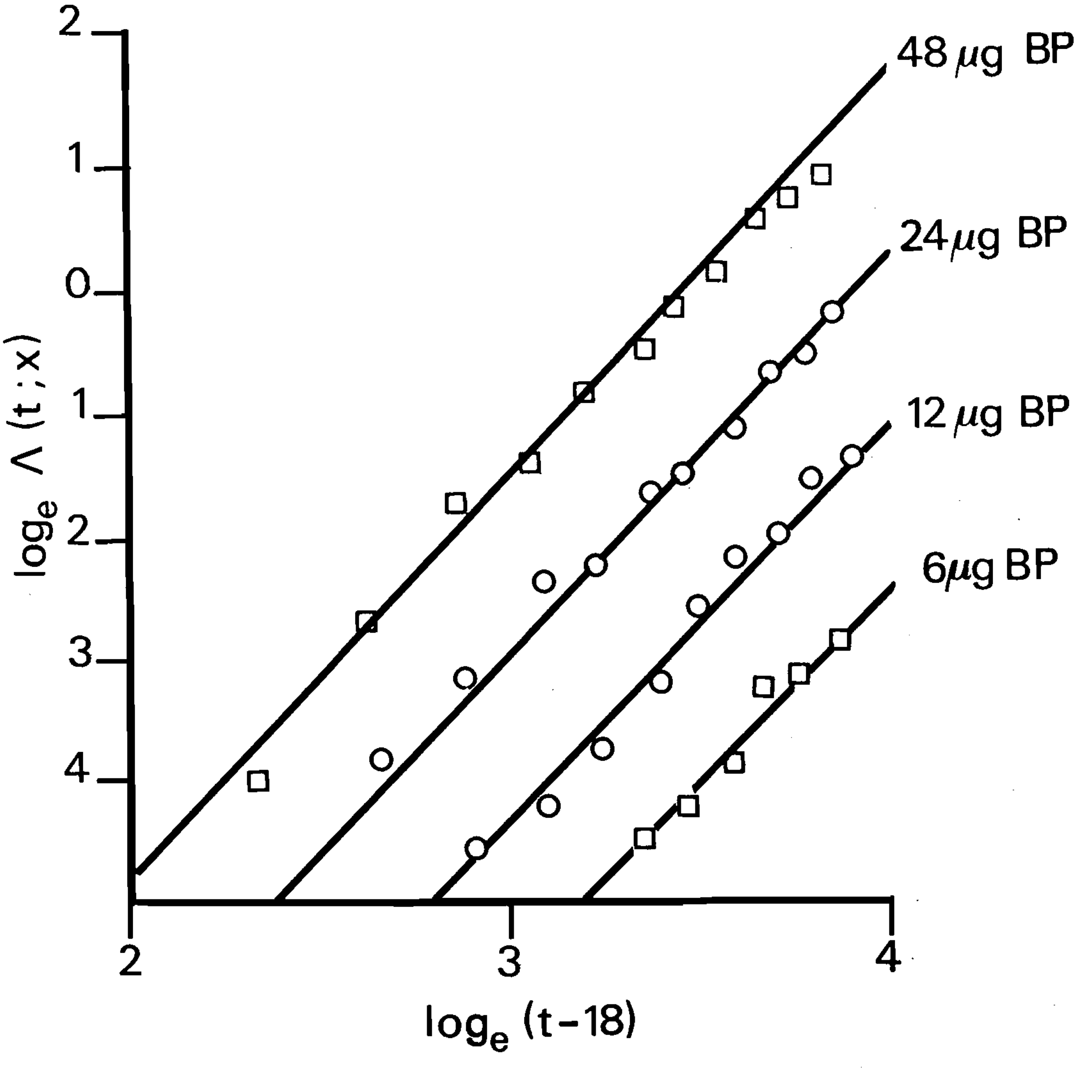
i.e.,

$$\Lambda(t;x) = Cx^{1.78}(t-18)^{2.95},$$

where C is a constant. It follows that the ratios of cumulative incidence rates for successive dosage groups $\Lambda(t;2x) \div \Lambda(t;x)$ and hence the ratios of the age-specific rates $\lambda(t;2x) \div \lambda(t;x)$, are equal to $2^{1.78} = 3.41$. Thus, within the range of dosage and ages of animals considered in this experiment, the effect of BP on incidence can be described very simply: a doubling of dose will lead to an approximate 3.4-fold increase in the age-specific skin tumour incidence rates.

The same investigators have shown in later work (Peto et al., 1975) that the relevant time variable is in fact not the age of the animal, but rather the duration of exposure to BP. They also point out that, since the powers of dose x and time t in the fitted formula for cumulative incidence are roughly 2 and 3,

Fig. 6.6 Estimated cumulative incidence rates of skin tumours occurring among female albino mice given weekly paintings of benzo[*a*]pyrene at four dosages, with parallel regression lines fitted by maximum likelihood (from Lee & O'Neill, 1971)



respectively, the data are consistent with a multi-stage theory for the origin of cancer wherein two of the three stages are affected by the carcinogen (Peto, 1977). Recent data for cigarette smoking and lung cancer in the British doctor study likewise suggest that incidence is proportional to the square of the dose rate (Doll & Peto, 1978).

If the linear logistic model were to be used to represent the data in the above experiment, this would take the form

$$\text{logit } P_i(x) = \alpha_i + \beta \log(x)$$

for the probability that an animal treated with x units of BP who is still at risk at age t_i develops a skin tumour within that week. The α_i parameters in turn could be modelled $\alpha_i = \alpha + \gamma \log(t_i)$ as a linear function of log age. There is no problem here with the fact that the logarithm of a zero dose is $-\infty$ and thus the estimated probability of tumour development 0, since skin tumours do not appear spontaneously on the backs of mice without treatment. For other studies, especially with humans, one could substitute a dose metameter of the form $z = \log(x + x_0)$, where x was the measured dose while x_0 represented a small *background dose* which was presumably responsible for any spontaneous cases. Although in principle it is possible to estimate x_0 from the data by maximum likelihood, this is rarely done. Special programmes would be required for such estimation since x_0 does not enter the regression equation in the same linear fashion as the other parameters. Furthermore, since different combinations of x_0 and β can give virtually identical fits to the data, the standard errors and covariances for the jointly estimated parameters tend to be large. Hence the best practice may simply be to assign x_0 some small value on the basis of *a priori* considerations. With the Ille-et-Vilaine tobacco data, we set $x_0 = 1$ and noticed that the resulting curve seemed to fit the observed data reasonably well (Figure 6.5).

6.12 Studies of interaction in a series of 2 × 2 tables

One of the principal advantages of using the logistic regression model is that it encourages quantitative description of how the changes in risk associated with one factor are modified by the interaction effects of other risk or nuisance variables. Since the Ille-et-Vilaine data are notably lacking in such interactions, they cannot be used to illustrate this important feature of statistical modelling. Hence in this section we analyse another set of published data, which happen to be in the form of a series of 2 × 2 tables, for which strong interaction effects are present.

Presence of interaction effects in a series of 2 × 2 tables means that the odds ratios depend systematically on the variables used for strata formation. Such dependence may have important implications for the nature of the disease process. The data we shall consider are those of Stewart and Kneale (1970) who hypothesized that the distribution of age at diagnosis for childhood cancers caused by obstetric X-rays was more concentrated or "peaked" than the age distribution of idiopathic childhood cancers. If this were so the risk ratio for irradiated *versus* non-irradiated children would also show a peak when plotted against age. The effect would presumably occur because the time of exposure for the radiogenic cases is limited to the period of gestation, while for other cancers it could vary over a broader age span.

Such variations are detected by the addition of interaction terms to the logistic model. In § 6.5 we considered a model in which the log relative risk was assumed to change linearly over the six age strata. More generally one might define several different regression variables, including transformations and cross-product terms, from factors such as age and time which are used to define strata. Let us denote by z_{il} the value of the l^{th} variable for the i^{th} stratum ($i = 1, \ldots, I$; $l = 1, \ldots, L$). Then the interaction model may be written

$$\text{logit } P_i(x) = \alpha_i + \beta x + \sum_{l=1}^{L} \gamma_l x z_{il}, \tag{6.25}$$

where as usual $P_i(x)$ denotes the disease probability in the i^{th} stratum for an exposed ($x = 1$) or unexposed ($x = 0$) individual. A consequence of this formulation is that the log relative risk for the i^{th} stratum is expressed

$$\log \psi = \log \frac{P_i(1)Q_i(0)}{P_i(0)Q_i(1)} = \beta + \sum_{l=1}^{L} \gamma_l z_{il}$$

as a linear function of the regression variables z, with the "constant" term β denoting the baseline log relative risk for the group having covariate values $\mathbf{z} = \mathbf{0}$. It is best to code the covariates in such a way that $\mathbf{z} = \mathbf{0}$ corresponds to some "typical" individual.

Summary data from the Oxford Childhood Cancer Survey and associated studies reported by Kneale (1971) are presented in Appendix II. Cases were ascertained as all children under ten years of age in England and Wales who died of cancer (leukaemia or solid tumours) during the period 1954–65. For each of these a neighbourhood control of the same age was selected who was alive and well at the time the case died. Only "traced" pairs, for whom both case and control mothers could be found and interviewed, were analysed. The published data ignore the exact pairing but do preserve the stratification by age and year of birth.

Exposure in this example is simply a question of whether or not the study subjects received *in utero* irradiation, as reported by the mother. The stratification variables were age at death, from 0 to 9 years, and year of birth, from 1944 to 1964. Because of the limited period of case ascertainment, not all 210 possible combinations of these factors appear. For example, among childhood cancer patients born in 1944, only those who died at age 9 are represented. A total of 120 such strata were available.

In order to estimate the overall relative risk of obstetric radiation, and to determine whether, and if so how, it varied with age and year, we fitted several versions of the model (6.25). Five different regression variables were used: z_1 = year of birth, coded $z_1 = -10$ for 1944, ..., $z_1 = 10$ for 1964; $z_2 = z_1^2 - 22$; z_3 = age at death, coded -9 for age 0, -7 for age 1, ..., 9 for age 9; $z_4 = z_3^2 - 33$; and $z_5 = z_1 \times z_3$. Different subsets of these were entered into the regression equation so as to detect particular kinds of trends and patterns in the relative risk.

Results of the analysis are shown in Table 6.16. Degrees of freedom (DF) for each model were obtained in the usual manner by subtracting the number of parameters, in this case the 120 α's plus additional β and γ terms, from the number of binomial observations, namely 240. The first model, which includes only the α's, assumes that the relative risk is unity in each stratum. In view of the large goodness-of-fit statistics, this supposition is clearly untenable. The second model specifies a constant relative

Table 6.16 Results of fitting several logistic regression models with interactions: Oxford study of obstetric radiation and childhood cancer[a]

Model	No. of parameters	DF	Goodness-of-fit statistics		Regression coefficients ± S.E.					
					Log RR	Interactions with				
			G	$\tilde{G}$		YR[b]	$YR^2 - 22$[c]	AGE[d]	$AGE^2 - 33$[c]	YR × AGE
					($\hat{\beta}$)	($\hat{\gamma}_1$)	($\hat{\gamma}_2$)	($\hat{\gamma}_3$)	($\hat{\gamma}_4$)	($\hat{\gamma}_5$)
1	0	120	207.89	196.74						
2	1	119	124.29	118.75	0.5102±0.0564					
3	2	118	116.96	112.52	0.5218±0.0567	–0.0390±0.0145				
4	3	117	111.57	108.74	0.5707±0.0611	–0.0450±0.0150	0.0068±0.0030			
5	3	117	116.33	112.75	0.5297±0.0576	–0.0312±0.0176		0.0105 ± 0.0133		
6	6	114	110.20	107.32	0.4738±0.1308	–0.0411±0.0182	0.0029±0.0057	0.0069±0.0134	0.0025±0.0028	–0.0054±0.0063

[a] From Breslow (1976); data from Kneale (1971)

[b] YR is coded as follows: 1944 = –10, 1945 = –9, . . ., 1963 = 9, 1964 = 10.

[c] Constants subtracted from square of AGE and YR so that variables sum to zero over tables

[d] AGE is coded as follows: 9 years = 9, 8 years = 7, 7 years = 5, . . ., 1 year = –7, 0 year = –9.

risk for obstetric radiation, estimated as $\hat{\psi} = \exp(0.5102) = 1.67$. Since the chi-square statistics for it are close to their mean values (DF), they might be taken as evidence of a good fit. However the introduction of a linear interaction term in year of birth (Model 3) results in a significant improvement ($G_2 - G_3 = 124.29 - 116.96 = 7.3$, $p = 0.007$). Hence there is reasonably strong evidence for a decrease in relative risk with year of birth. Additional improvement in fit occurs when a quadratic term in year is added to the model, which would indicate a degree of curvature in the regression line. However it is of lesser statistical significance ($G_3 - G_4 = 5.39$, $p = 0.02$). Figure 6.7 shows age-adjusted estimates of the log-relative risk for each year, together with linear and quadratic regression lines as fitted by Models 3 and 4. This illustrates graphically the nature of the decline in the radiation effect over time.

Absolutely no improvements in fit accompanied the addition to the model of either linear or quadratic terms in age: compare Models 3 *versus* 5 and 4 *versus* 6. The quadratic term would be expected to be particularly sensitive to a peak in relative risk as a function of age. The lack of evidence for any such peak argues against the hypothesis that the age distributions for radiogenic and idiopathic cancers are different. Improvements in radiological technology probably account for the declining effect with year of birth (Bithel & Stewart, 1975).

Fig. 6.7 Age-adjusted estimates of log relative risk (odds ratio) for obstetric radiation each with approximate 80 percent confidence limits and both linear and quadratic regression lines (from Breslow, 1976; data from Kneale, 1971)

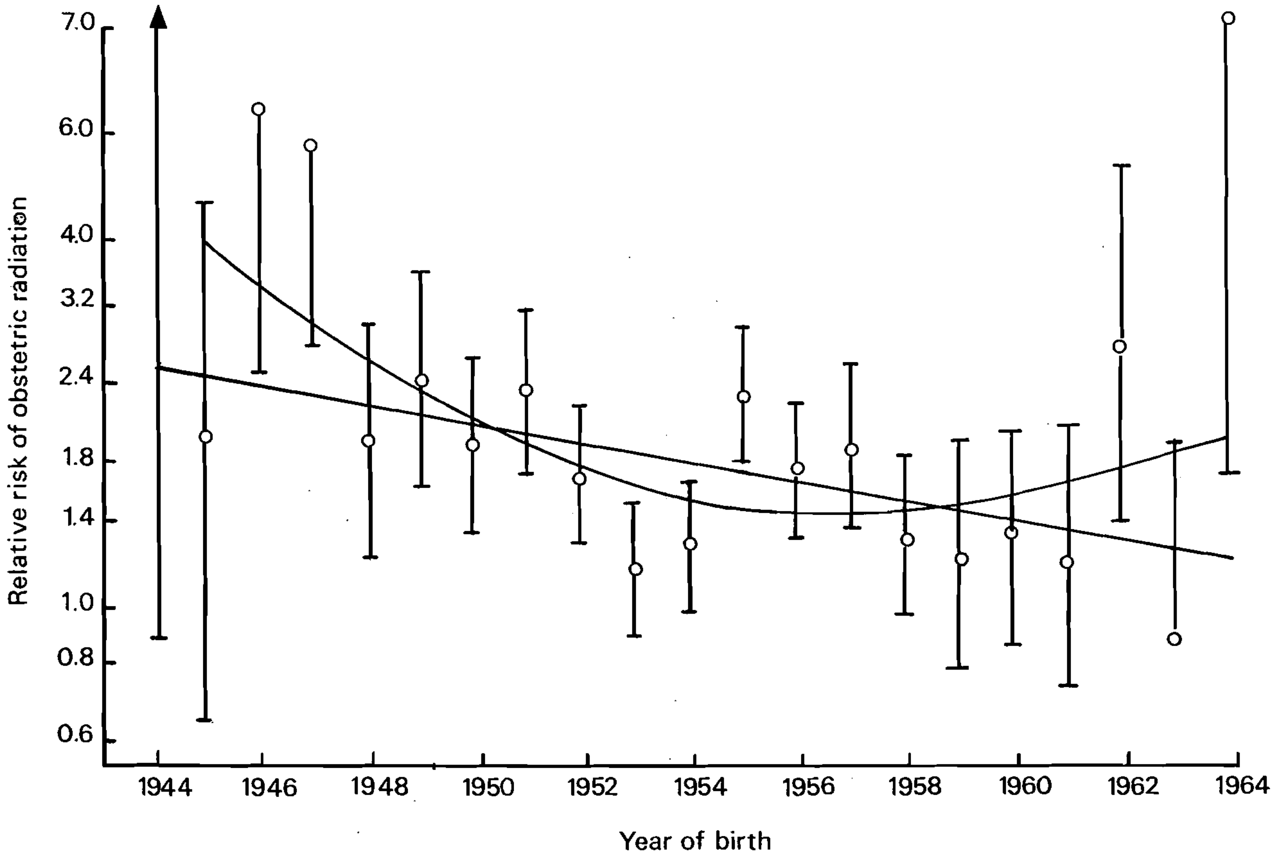


Table 6.17 Comparison of the log relative risk and its interaction with year of birth, depending on the degree of polynomial adjustment for age and year: Oxford study of obstetric radiation and childhood cancer[a]

Degree of polynomial in age and year	No. of parameters	DF	Goodness of fit $\tilde{G}$	Estimates of log-relative risk and interaction $\hat{\beta} \pm$ S.E.	$\hat{\gamma}_1 \pm$ S.E.
0	3	237	114.62	0.5113 ± 0.0562	−0.0343 ± 0.0136
1	5	235	113.87	0.5124 ± 0.0562	−0.0382 ± 0.0143
2	8	232	113.57	0.5124 ± 0.0562	−0.0384 ± 0.0143
3	12	228	113.55	0.5150 ± 0.0563	−0.0385 ± 0.0143
4	17	223	113.43	0.5157 ± 0.0564	−0.0385 ± 0.0143
5	23	217	113.33	0.5163 ± 0.0564	−0.0386 ± 0.0144
Stratified[b]	122	118	112.52	0.5218 ± 0.0567	−0.0390 ± 0.0145

[a] From Breslow and Powers (1978); data from Kneale (1971)
[b] From Model 3, Table 6.16

As shown in § 6.8, an alternative to a stratified analysis is simply to model the effects of the nuisance factors on disease incidence, replacing the α_i in the logistic model by quantitative terms. In order to compare the results from such an analysis with those just obtained, we considered analogs of Model 3 in which the log-relative risk was assumed to decline linearly with year of birth. Polynomials of increasing degree in age and year were used to give different degrees of adjustment for the confounding effects of these factors. Thus the models fitted were of the form

$$\text{logit}\, P_i(x) = \begin{cases} \alpha_0 + \beta x + \gamma_1 x z_{i1} & \text{(unadjusted)} \\ \alpha_0 + \alpha_1 z_{i1} + \alpha_3 z_{i3} + \beta x + \gamma_1 x z_{i1} & \text{(linear)} \\ \alpha_0 + \sum_1^5 \alpha_l z_{il} + \beta x + \gamma_1 x z_{i1} & \text{(quadratic)} \end{cases}$$

and so on, using third, fourth and fifth degree polynomials. The results in Table 6.17 show that increasing the degree of polynomial adjustment leads to better agreement with results of the stratified analysis (Breslow & Powers, 1978). It is somewhat surprising that there is so little improvement in the fit, and so little change in the estimated relative risks, as more terms of age and year are included. A partial explanation is, of course, that the sample was deliberately stratified to ensure that the numbers of cases and controls in each age/year stratum were equal. Thus one might not expect these two factors to contribute significantly to a model designed to discriminate cases from controls. However, as discussed in § 3.4, this identity of the marginal distributions of age and year for cases and controls is not sufficient to justify ignoring these factors in the analysis. *In general, variables used for stratification or matching in the design stage must also be accounted for in the analysis in order to obtain unbiased estimates of the relative risk.* An example which better illustrates this point is presented in § 7.6. If strata are formed at the time of analysis, rather than by design, there will be imbalances in the numbers of cases and controls within strata, and the differences between the stratified and unadjusted analyses will be more obvious than they are in Table 6.17.

REFERENCES

Anderson, J.A. (1972) Separate sample logistic discrimination. *Biometrika, 59,* 19–35

Armitage, P. (1975) *Statistical Methods in Medical Research,* Oxford, Blackwell

Baker, R.J. & Nelder, J.A. (1978) *The GLIM System. Release 3,* Oxford, Numerical Algorithms Group

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice,* Cambridge, Mass., MIT Press

Bithel, J. & Stewart, A. (1975) Pre-natal irradiation and childhood malignancy: a review of British data from the Oxford Survey. *Br. J. Cancer, 31,* 271–287

Breslow, N.E. (1975) Analysis of survival data under the proportional hazards model. *Rev. Int. Stat., 43,* 45–58

Breslow, N.E. (1976) Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics, 32,* 409–416

Breslow, N.E. (1978) The proportional hazards model: applications in epidemiology. *Comm. Stat.-Theor. Meth., A7,* 315–332

Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.L. & Sabai, C. (1978) Estimation of multiple relative risk functions in matched case-control studies. *Am. J. Epidemiol., 108,* 299–307

Breslow, N.E. & Powers, W. (1978) Are there two logistic regressions for retrospective studies? *Biometrics, 34,* 100–105

Cornfield, J. (1962) Joint dependence of the risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Fed. Proc., 21,* 58–61

Cox, D.R. (1970) *The Analysis of Binary Data,* London, Methuen

Cox, D.R. (1972) Regression models and life tables (with discussion). *J. R. stat. Soc. B, 34,* 187–220

Cox, D.R. & Hinkley, D.V. (1974) *Theoretical Statistics,* London, Chapman & Hall

Day, N.E. & Byar, D.P. (1979) Testing hypotheses in case-control studies – equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics, 35,* 623–630

Day, N.E. & Kerridge, D.F. (1967) A general maximum likelihood discriminant. *Biometrics, 23,* 313–323

Doll, R. & Peto, R. (1978) Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers. *J. Epidemiol. Community Health, 32,* 303–313

Efron, B. (1975) The efficiency of logistic regression compared to normal theory discriminant analysis. *J. Am. stat. Assoc., 70,* 892–898

Farewell, V.T. (1979) Some results on the estimation of logistic models based on retrospective data. *Biometrika, 66,* 27–32

Fienberg, S.E. (1977) *The Analysis of Cross-Classified Data,* Cambridge, Mass., MIT Press

Haberman, S.J. (1974) *The Analysis of Frequency Data,* Chicago, University of Chicago Press

Halperin, M., Blackwelder, W.C. & Verter, J.I. (1971) Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. *J. chron. Dis., 24,* 125–158

Kneale, G.W. (1971) Problems arising in estimating from retrospective survey data the latent period of juvenile cancers initiated by obstetric radiography. *Biometrics, 27,* 563–590

Lee, P. & O'Neill, J. (1971) The effect of both time and dose on tumour incidence rate in benzopyrene skin painting experiments. *Br. J. Cancer, 25,* 759–770

Lininger, L., Gail, M., Green, S. & Byar, D. (1979) Comparison of four tests for equality of survival curves in the presence of stratification and censoring. *Biometrika, 63,* 419–428

Mantel, N. (1973) Synthetic retrospective studies and related topics. *Biometrics, 29,* 479–486

Mosteller, F. & Tukey, J.W. (1977) *Data Analysis and Regression: a Second Course in Statistics,* Reading, Mass., Addison & Wesley

Nelder, J.A. (1977) A reformulation of linear models. *J. R. stat. Soc. A, 140,* 48–77

Peto, R. (1977) *Epidemiology, multistage models and short-term mutagenicity tests.* In: Hiatt, H.H., Watson, J.D. & Winsten, J.A., eds. *Origins of Human Cancer,* Cold Spring Harbor, NY, Cold Spring Harbor Publications, Vol. 4, pp. 1403–1421

Peto, R., Roe, F., Lee, P., Levy, L. & Clark, J. (1975) Cancer and ageing in mice and men. *Br. J. Cancer, 32,* 411–426

Prentice, R.L. & Breslow, N.E. (1978) Retrospective studies and failure time models. *Biometrika, 65,* 153–158

Prentice, R.L. & Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika, 66,* 403–411

Press, S.J. & Wilson, S. (1978) Choosing between logistic regression and discriminant analysis. *J. Am. stat. Assoc., 70,* 699–705

Rao, C.R. (1965) *Linear Statistical Inference and its Applications,* New York, Wiley

Seigel, D.G. & Greenhouse, S.W. (1973) Multiple relative risk functions in case-control studies. *Am. J. Epidemiol., 97,* 324–331

Stewart, A. & Kneale, G.W. (1970) Age-distribution of cancers caused by obstetric x-rays and their relevance to cancer latent periods. *Lancet, ii,* 4–8

Thomson, W.A., Jr (1977) On the treatment of grouped observations in life studies *Biometrics, 33,* 463–470

Truett, J., Cornfield, J. & Kannel, W. (1967) A multivariate analysis of the risk of coronary heart disease in Framingham. *J. chron. Dis., 20,* 511–524

Vitaliano, P.O. (1978) The use of logistic regression for modelling risk factors: with application to non-melanoma skin cancer. *Am. J. Epidemiol., 108,* 402–414

Walker, S.H. & Duncan, D.B. (1967) Estimation of the probability of an event as a function of several independent variables. *Biometrika, 54,* 167–179

LIST OF SYMBOLS – CHAPTER 6 (in order of appearance)

logit(p)	the logistic transform of a proportion p; $\log(\frac{p}{1-p})$
x	risk variable
ψ	odds ratio
β	log odds ratio

P_1 disease probability for exposed

P_0 disease probability for unexposed

$P(x)$ disease probability for exposure to an amount x

$r(x)$ relative risk of disease associated with exposure to an amount x

α log odds for disease among unexposed

P_{ij} disease probability associated with exposure to level i of factor A and level j of factor B

Q_{ij} $1\text{-}P_{ij}$

ψ_{ij} odds ratio associated with exposure to level i of factor A and level j of factor B ($\psi_{00} = 1$)

r_A relative risk of exposure to factor A

r_B relative risk of exposure to factor B

r_{AB} relative risk of exposure to both factor A and factor B

β_1 log odds ratio associated with exposure to factor A

β_2 log odds ratio associated with exposure to factor B

γ (multiplicative) interaction parameter; log of the ratio of the relative risk for combined exposure divided by the product of relative risks for individual exposures

$P(x_1,x_2)$ disease probability associated with exposure to an amount x_1 of factor A and x_2 of factor B

β_k coefficient of variable x_k in logistic regression equation; log relative risk associated with unit increase in x_k

γ_{kl} coefficient of cross product variable $x_k x_l$ in logistic regression equation; interaction parameter

P_{ijk} disease probability associated with exposure to levels i of A, j of B and k of C

Q_{ijk} $1\text{-}P_{ijk}$

ψ_{ijk} relative risk associated with exposures to levels i of A, j of B and k of C

γ_{ijk} coefficient of variable $x_i x_j x_k$ in logistic regression equation; second order interaction parameter

$\mathrm{pr}(\ |\ )$ probability of one event given the occurrence of another

i subscript indicating one of I strata

$\mathbf{x}$ vector of risk variables associated with an individual

$P_i(\mathbf{x})$ disease probability associated with a vector $\mathbf{x}$ of risk variables in the i^{th} stratum of the population

y binary response variable; $y = 1$ for diseased, $y = 0$ for disease-free

π_1 sampling fraction for cases; probability that a diseased person is included in the study as a case

π_0 sampling fraction for controls; probability that a disease-free person is included in the study as a control

z indicator sampling variable: $z = 1$ for inclusion in the study, $z = 0$ otherwise

$\boldsymbol{\Sigma}$ covariance matrix for the distribution of risk variables, assumed common for cases and controls

$\underset{\sim}{\boldsymbol{\mu}}_1$ expected values of risk variables among cases

$\underset{\sim}{\boldsymbol{\mu}}_0$ expected values of risk variables among controls

$\bar{\mathbf{x}}_1$ sample mean of risk variables $\mathbf{x}$ among cases

$\bar{\mathbf{x}}_0$ sample mean of risk variables $\mathbf{x}$ among controls

$\mathbf{S}_p^2$ covariance matrix of risk variables pooled from separate samples of cases and controls

$\mathbf{l}$ denotes a partition of the integers from 1 to n into two groups, one of size n_1 and the other of size $n_0 = n-n_1$; e.g., if $n_1 = 2$ and $n_0 = 3$ a possible partition is $l_1 = 3, l_2 = 4, l_3 = 1, l_4 = 2, l_5 = 5$ or $\mathbf{l} = (3,4,1,2,5)$

$\mathbf{x}_j$ vector of risk variables for jth study subject

G goodness-of-fit statistic based on the log likelihood

$\mathbf{S}$ efficient score; vector first of first derivatives of the log likelihood function

$\mathbf{I}$ information matrix; matrix of negatives of second partial derivatives of the log likelihood function

Z standardized regression coefficient (equivalent normal deviate)

$\tilde{G}$ chi-square goodness-of-fit statistic for grouped data, based on differences between observed and expected values

(N.B. Subscripts on the above quantities G, $\tilde{G}$, $\mathbf{S}$, $\mathbf{I}$, and Z denote their values under different models)

O observed number of cases (or controls) in a particular cell with grouped data

E expected number of cases (or controls) in a cell, predicted by fitted model

ϱ correlation coefficient between two variables

$\Lambda(t;x)$ cumulative incidence of skin tumours by week t among animals continuously exposed to BP at a dose rate x

$\lambda(t;x)$ age-specific incidence of skin tumours at week t among animals continuously exposed to BP at a dose rate x

L number of regression variables (covariates) associated with each of a series of 2 × 2 tables

l lth of L covariates associated with a series of 2 × 2 tables

z_{il} value of the lth covariate for the ith of a series of 2 × 2 tables

^ when placed over another symbol this indicates an estimate of a population parameter calculated from the sampled data; or a fitted cell frequency predicted from a model; e.g., $\hat{\beta}$ is an estimate of β