NAZIOARTEKO ESTATISTIKA
MINTEGIA EUSKADIN

1989

SEMINARIO INTERNACIONAL
DE ESTADISTICA EN EUSKADI

GENERALIZED LINEAR MODELS IN EPIDEMIOLOGY

J. C. DUFFY

EUSKAL ESTATISTIKA - ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

EUSTAT

NAZIOARTEKO ESTATISTIKA
MINTEGIA EUSKADIN
1989
SEMINARIO INTERNACIONAL
DE ESTADISTICA EN EUSKADI

GENERALIZED LINEAR MODELS IN EPIDEMIOLOGY

J. C. DUFFY

KOADERNOA 17 CUADERNO

EUSTAT

Lanketa / *Elaboración:*

Euskal Estatistika-Erakundea /
Instituto Vasco de Estadística

Argitalpena / *Edición:*

Euskal Estatistika-Erakundea /
Instituto Vasco de Estadística
C/ Dato 14-16 - 01005 Vitoria-Gasteiz

Botaldia / *Tirada:*

1.000 ejemplares
I-1989

Fotokonposaketarako tratamendu informatikoa:
Tratamiento informático de fotocomposición:

Fotocomposición IPAR, S.C.L.
Particular de Zurbarán, 2-4 - 48007 BILBAO

Inprimaketa eta Koadernaketa:
Impresión y Encuadernación:

ITXAROPENA, S.A.
Araba kalea, 45 - Zarautz (Gipuzkoa)

Lege-gordailua / *Depósito legal:* S.S. 57/90

ISBN: 84-7542-127-10 Obra completa
ISBN: 84-7749-062-7

JOHN C. DUFFY. BIOGRAPHICAL OUTLINE

Lecturer in statistics, University of Edinburgh and research scientist, M.R.C. unit for epidemiological studies in psychiatry. Research interests and publications in the fields of epidemiology, causal modelling, survey methodology. Particular areas of expertise include epidemiological applications in cataract, depression, alcohol-related problems and suicide. Past member of RSS social statistics section committee, president of Association of University Teachers (Scotland), joint prizewinner, best paper of 1984, study group for the use of computers in survey analysis.

GENERALIZED LINEAR MODELS IN EPIDEMIOLOGY: TOPICS OF THE COURSE

Log-linear and logistic-linear modelling; GLIM computer package; odds ratio and relative risk; cohort studies; case-control studies; stratification; Mantel-Haenszel estimates; matched studies; analysis of vital rates and age-period-cohort models; survival methods.

CONTENTS

I. COMPARATIVE STUDIES OF FREQUENCIES 9

1.1. General population studies 12
1.2. Relative measures of disease incidence 13
1.3. Cohort studies 14
1.4. Case-control studies 15
1.5. Matched case-control studies 16
1.6. The logistic-linear model 17
1.7. GLIM example 1 18
1.8. Interval estimation 20
1.9. Extension to case-control studies 22

II. MULTIVARIATE STUDIES 25

2.1. Confounding 27
2.2. Confounding variables at several levels 31
2.3. Independent variables, variates or factors? 36
2.4. Analysis of ungrouped data 40
2.5. Analysis of matched pairs 43
2.6. Multivariate modelling and joint action 46
2.7. Analysis of vital rates 47
2.8. Age-period-cohort models 48
2.9. A note on the identification problem 51

III. SURVIVAL ANALYSIS IN GLIM 53

3.1. Basic concepts 55
3.2. Introductory theory 55
3.3. Lifetables and the Kaplan-Meier survival estimate 57
3.4. The Logrank test 62
3.5. Cox regression and the proportional hazards model 63

BIBLIOGRAPHY 67

APPENDIX A: Drinking days, weekly consumption and experience of alcohol dependence symptoms of 223 company directors.......... 69

APPENDIX B: Appendectomy data-case-control differences.......... 71

APPENDIX C: Survival of adendid cystic carcinoma patients 72

APPENDIX D: Parasuicide repetition data 73

I. COMPARATIVE STUDIES OF FREQUENCIES

Generalized Linear Models in Epidemiology

Many epidemiological investigations involve the collection and analysis of counted data, such as *frequencies* of occurrence of events, numbers of individuals having particular attributes and so on. Also of importance are studies of *times* of occurrence of events, for example times of death, times to relapse or times of onset of a particular illness or other condition.

Methods of analysing data of these kinds have advanced greatly since the development of Generalized Linear Model theory, and have been facilitated by implementation of the methodology on computers, initially through the GLIM package, and now on most of the generally available major statistical analysis systems, such as SPSS-X, GENSTAT, SAS and BMDP. The flexibility of model formulation and implementation (particularly on the GLIM package) permits its use for the analysis of many different types of epidemiological investigation. This text will consider three main categories of study, the comparative study of frequencies, survival analysis and the analysis and modelling of vital statistics.

It will be assumed that the reader is familiar with elementary methods of analysing frequency data, such as the χ^2 test, the exact test for 2x2 contingency tables, McNemar's test for paired data and with methods of multiple regression analysis. Some knowledge of the main tools of epidemiological investigation and their purposes will be taken for granted.

The intention of the present work is that the reader will be able to understand the logic of applications of the General Linear Model in epidemiology, will gain some facility in identifying appropriate methods of analysis for certain types of epidemiological investigation and will be able to undertake some analyses themselves using the GLIM computer package. Detailed consideration of statistical methodology is impossible given the aims of the course and the short time available, but Breslow & Day (1980) cover theoretical aspects of logistic-linear modelling applied to epidemiology in some depth and McCullagh & Nelder (1983) is a key reference to generalized linear model theory.

1. Comparative studies of frequencies

In this section I will consider three major types or categories of study. These are general population studies, cohort studies and case-control studies. To introduce and motivate consideration of the problems posed by each of these types it is necessary to discuss some statistical quantities of particular relevance to epidemiological studies.

1.1 General population studies

One method of investigating the association between the occurrence of an illness and exposure to possible risk factors is to select a random group of individuals from the population for investigation. These subjects may be assessed for presence of the illness and a number of risk factors at the time of investigation (a cross-sectional study), or may be followed up for a number of years in a longitudinal or prospective study. The presence or development of the illness in question in the individuals is recorded, as are their attributes (potential risk factors). In a study of this type the outcome of interest, that is the development of illness, may be considered as a binary variable, in the case in which the total number of individuals with the illness is analysed, or as a time, when we analyse the times to development of the illness. The methods of this section are appropriate to the former case.

If we follow up a random sample from the appropriate population there is one basic statistical parameter which we can estimate, and that is the incidence rate of the illness, defined simply as the number of individuals developing the illness divided by the total number in the study. An annual rate may be calculated by dividing this quantity by the time period of the study in years . It is common also to express rates as rates per 100,000 or per million, which simply involves multiplication by the appropriate figure. Rates should be corrected if subjects leave the study either through death, cessation of cooperation or for any other reason. The denominator after such correction becomes person years at risk, calculated appropriately. It is also useful in some situations to estimate the prevalence rate, which is simply the proportion of individuals with the illness in question at a single point in time.

In a cross-sectional study only the prevalence rate can be estimated. This poses certain difficulties, since the relationship between prevalence and incidence is rather complicated, involving such factors as duration of illness, time to death, and incidence of cure. Nevertheless, prevalence may be analysed by the same methods as incidence, but the interpretation of such analyses requires some caution.

The purpose of any epidemiological study is to examine the effect of possible risk factors on disease. Considering the case of the longitudinal study further, we might classify individuals in the study on the presence or absence of a particular attribute, a potential risk factor. We may then estimate incidence rates for those with and without the risk factor.

Table 1

		Disease		
		Present	Absent	Total
Risk	Present	a	b	m_1
Factor	Absent	c	d	m_2
	Total	n_1	n_2	n

Table 1 shows how the information from such a classification might be presented. Entries in the table are frequencies corresponding to each category. It is important to notice that the only quantity in the table which is fixed in advance is the total number of individuals in the study, n. The population rate of illness may be estimated as n_1/n, and the rates in the risk factor present and absent groups are estimated as a/m_1 and c/m_2 respectively. A test of significance of the difference between these may be carried out by means of a χ^2 test, or equivalently a test for the difference between two proportions, unless n is particularly small, in which case Fisher's exact test is appropriate. Confidence intervals are easily constructed for the difference in proportions by standard methods, either approximately from the normal distribution or using the more accurate quadratic bounds, or for small n the hypergeometric distribution.

However the difference between the two rates is not the standard measure of association between risk factor and disease. There are a number of reasons for this, and we will see later that many common types of epidemiological study do not permit estimation of the rate difference. The measures in most common use are the **relative risk** and **odds ratio.**

1.2 Relative measures of disease incidence

The relative risk of disease associated with a particular risk factor is defined as the ratio of the rate among those exposed to the risk factor to the rate among the unexposed. In terms of table 1 above the relative risk is estimated by

$$\hat{rr} = \frac{a/m_1}{c/m_2} \qquad (1.2.1)$$

This quantity measures how much more or less likely it is that disease occurs among those exposed to the factor in question. It takes values between 0 and infinity, and represents positive or negative association between illness and exposure according as rr is greater or less than one respectively.

The odds ratio is another common measure of association. The probability of an event e can be re-expressed in terms of odds, O(e) as follows

$$O(e) = \frac{p}{1-p} \qquad (1.2.2)$$

where p is the probability of the event in question. Thus, given rates of illness, which may be considered as probabilities of illness we may calculate the odds for the exposed and unexposed groups.

In terms of the quantities in table 1 the odds ratio ψ may be estimated in a simple form as

$$\hat{\psi} = \frac{ad}{bc} \qquad (1.2.3)$$

The most attractive feature of the odds ratio as a measure of relative association is that it is estimable in many situations where the relative risk is not, as we shall see in the next sections. Another useful aspect is that it approximates the relative risk quite well if the incidence rate of the disease is small.

To see this consider an illness with rates r_e and r_u among the exposed and unexposed members of the population. The relative risk is then

$$rr = \frac{r_e}{r_u} \qquad (1.2.4)$$

and the odds ratio is

$$\psi = \frac{r_e(1-r_u)}{r_u(1-r_e)} \quad (1.2.5)$$

If r_u and r_e are small then clearly the quotient of the two bracketed terms is approximately 1.

As an example of a prospective study consider the following data from a longitudinal study of British civil servants (Marmot et al, 1981). The outcome of interest is death during the ten years of the study, and the table further classifies the subjects on the basis of their alcohol consumption. In the original article these data were presented in a more detailed classification, which we shall consider later.

Table 2: Deaths in a 10-year study of English civil servants

		Outcome		
		Death	Survival	Total
Alcohol	Zero	45	432	477
Consumption	>0	68	877	945
Total		113	1309	1422

From table 2 we can easily calculate the 10-year death rates as 0.0943 among the non-drinkers and 0.0720 among the drinkers. The estimate of relative risk is thus 1.31 considering the non-drinkers as the exposed category. The odds ratio estimate is 1.34 showing a reasonably close approximation to the relative risk. The usual χ^2 test may be applied to the table and yields a value of 2.17 on 1 degree of freedom, which does not indicate a significant association between alcohol and mortality.

1.3 Cohort studies

One way of assessing the influence of a risk factor on mortality is to follow two groups over time, a group exposed to the risk factor, and the other group not exposed. The results of such a study can be tabulated and analysed by the methods of the previous section, but there is one difference worth noting. The population proportions exposed to the risk factor could be estimated from the type of general population study considered in the last section, but in a comparative cohort study the numbers of individuals in the exposed and unexposed groups, m_1 and m_2 in the notation of table 1, are determined by the experimenter.

However the rates of illness among the exposed and unexposed may be estimated from cohort studies, and an estimate of relative risk can be obtained directly. This is not so for the next type of study to be considered.

1.4 Case-control studies

In a case-control study the researcher compares a group of individuals with the disease in question (cases) with a group without the disease (controls). The numbers in these groups are determined by the investigator, and therefore such studies do not permit direct estimation of the rates of illness among the exposed or unexposed groups. To see this consider the results of a case-control study in the form of table 1.

Since n_1 and n_2 are determined by the investigator different choices of these will lead to different values of a/m_1 and c/m_2 . It follows that these cannot be estimates of the true rates among the exposed and unexposed, and hence the relative risk cannot be estimated (Cornfield, 1951). Knowledge of the population rate of illness is required in order to estimate the relàtive risk from a case-control study, and for many conditions such information is not available. However the odds ratio is estimable from such a study, and the estimate is just as in the previous case

$$\hat{\psi} = \frac{ad}{bc} \qquad (1.4.1)$$

The rationale of the estimate is slightly different. Essentially the argument is that the frequencies should be weighted to take account of the different sampling fractions for cases and controls, which would imply multiplying the entries in the disease present (case) column by a constant, X say, and the entries in the disease absent (control) column by Y. The odds for the exposed group may be estimated as the ratio of cases to controls in the exposed category, and similarly for the unexposed. This implies each set of odds is weighted by the same constant, X/Y. On taking the ratio of the two odds the constant cancels, giving the result. Another way of looking at this is to consider risk factor prevalence among the case and controls. If we let p_1 and p_2 be the proportions of cases and controls respectively with the risk factor, then it follows that our estimates of these would be a/n_1 and b/n_2 in the notation of table 1. Our estimates of the complementary probabilities, q_1 and q_2 are similarly c/n_1 and d/n_2. Thus the odds ratio may be expressed in terms of risk factor prevalences as p_1q_2/p_2q_1.

The data in Table 3 are taken from a case-control study of cataract in Edinburgh (Clayton et al, 1980). Notice that the population rate of cataract is certainly not 867/1190, these numbers being simply the sizes of the case and control groups. However the odds ratio for cataract due to smoking is estimable as

$$\hat{\psi} = \frac{298*236}{87*569} = 1.42$$

Table 3: Smoking and cataract in Edinburgh

	Cataract Cases	Population Controls	Total
Smoker	298	87	385
Non-smoker	569	236	805
Total	867	323	1190

The usual χ^2 test gives a value of 5.95 on 1 degree of freedom indicating strong evidence of association between smoking and cataract.

1.5 Matched case-control studies

Apart from the risk factor under investigation other variables such as age, sex and unknown risk factors may affect the probability of developing illness, and may also be associated with presence of the risk factor. Such variables are said to be *confounding* and one way to take account of known confounding factors is by individually matching each case to a control who shares the potentially confounding characteristics of the case. The resulting study design, the matched case-control study is rather different from the previous types considered, and data analysis is different also.

Table 4

		Control		
	Risk Factor	Present	Absent	Total
Case	Present	a	b	m_1
	Absent	c	d	m_2
	Total	n_1	n_2	n

The entries in table 4 are the frequencies of case-control *pairs* classified according to the presence or absence of the risk factor for each member of the pair. Thus for example there are b pairs in which the case member has the risk factor and the control does not. The total in the table, n, is simply the total number of pairs, which is of course one half the total number of individuals involved.

It is obvious that rates cannot be estimated directly from the data in table 4, and it might appear that the odds ratio will also pose problems of estimation. However the odds ratio is in fact estimated very easily as

$$\hat{\psi} = \frac{b}{c} \qquad (1.5.1)$$

This may be shown by considering the probabilities p_1 and p_2 that a case and a control are positive on risk factor exposure.

Defining $$q_1 = 1-p_1 \qquad (1.5.2)$$

and $$q_2 = 1-p_2 \qquad (1.5.3)$$

it follows that the probability that a discordant pair has the case member of the pair being exposed to the risk factor, and the control member not is given by

$$\frac{p_1q_2}{p_1q_2+q_1p_2} \qquad (1.5.4)$$

Similarly the probability of a discordant pair being such that the case member does not have the risk factor while the control member does is

$$\frac{p_2q_1}{p_1q_2+q_1p_2} \qquad (1.5.5)$$

Hence the estimate in equation (1.5.1) is the maximum likelihood estimate of the odds ratio, as the observed proportions are maximum likelihood estimates of the appropriate probabilities. Notice that pairs which are not discordant do not contribute to the estimation procedure, nor to statistical testing. The test used in these circumstances is a binomial test that the probability of each type of discordant pair is 0.5, given the total number of discordant pairs, which for large values of b+c may be performed by means of a χ^2 test on 1 degree of freedom. This procedure is often called McNemar's test.

1.6 The logistic-linear model

An alternative method of analysing studies of these types, and one which as we will see lends itself to more complicated analyses of higher dimensional tables, is provided by logistic regression, which is widely implemented in computer packages such as GLIM. This methodology may be introduced simply by a brief description of its theoretical basis and example analyses of the two sets of data considered in the last section.

The model considers a linear regression of log-odds of illness risk. For the exposed population the appropriate quantity is $\log(r_e/(1-r_e))$, and for the unexposed $\log(r_u/(1-r_u))$ in the notation of section 1.2. Notice that the difference of these two *logits* is simply $\log\psi$. Now consider a regression equation in which the dependent variable, y, is the logit of the rate of illness, and the independent variable, x , is exposure at levels 0 (corresponding to no exposure) and 1 (corresponding to exposure). The regression equation may be written

$$E[y] = \alpha+\beta x \qquad (1.6.1)$$

so for the unexposed group the model gives

$$\log(r_u/(1-r_u)) = \alpha \qquad (1.6.2)$$

and for the exposed group

$$\log(r_e/(1-r_e)) = \alpha+\beta \qquad (1.6.3)$$

Hence the parameter β in the model corresponds to the logarithm of the odds ratio due to exposure. Estimation of β and its standard error will permit point and interval estimation of the odds ratio, in addition to testing the significance of the effect of exposure.

The model may be fitted easily by maximum likelihood using the GLIM package. GLIM requires the probability distribution of the observations to be specified, and uses either an appropriate default LINK function, or one supplied by the user. The LINK function is that transformation of the expected value of the observations which is predicted by a linear function of the explanatory variates. In the case of binomial observations the logistic function is an appropriate LINK function and is automatically set by GLIM when the error structure is declared.

There are several important quantities calculated by the program, which are of value in various ways. GLIM will fit a null model, that is a model with no explanatory variables, which in the present context implies fitting the same rate to the exposed and unexposed groups. The *deviance* statistic for this model produced by the program (identified as residual deviance in the output) is the negative of twice the log-likelihood value at its maximum, that is, evaluated using the best single estimate of the rate of illness for both groups. Although the distribution of this statistic is not defined in the general case, in the particular instance of two groups and a binary explanatory variable it turns out to be the χ^2 distribution on 1 degree of freedom (as for the 2x2 table). This is because differences in deviance between models are interpretable as χ^2 statistics, and the full model for two groups involves only one parameter, β, the corresponding deviance having fitted this parameter being zero. The deviance statistic is not identical to the usual Pearson X^2 statistic, although the values will be very close, and for large n virtually equal.

1.7 GLIM example 1

Several examples of the use of GLIM will occur in this course, and while the directives used in these analyses will be explained, it is not the intention to give a complete overview of the GLIM package. Interested readers are referred to the GLIM introductory guide (Swan, 1985) and the GLIM manual (Payne, 1985) for a full description of the package, and to Healy (1988) for a general practical introduction.

For the present example the data in table 2 has been used, with exposure level 1 corresponding to the non-drinking group. In order to set up a GLIM analysis it is necessary to define the number of data points which determines the length of the arrays or vectors which will be named to contain the data. This is achieved through the $UNIts directive, and for the 2x2 contingency table the appropriate directive is

```
$UNITS 2
```

Notice that although there are 4 cells in the table, for logistic regression we actually only have 2 data points. This is because there are only two probabilities involved, as we condition on the marginal totals for the exposure categories. Note, too, that directive names are distinguished from other text by being preceded by a $ sign and that only the first three characters of a directive are required by the program, although for completeness they are often given in full in these examples. GLIM control statements may be entered in upper or lower case, but in order to distinguish them from other text they will be set in upper case throughout.

The next step is to define and then enter the data using the $DATa and $REAd directives as follows

```
$DATA EXP CASES N
$READ
1 45 477
2 68 945
```

The names of the data vectors are at choice, and in this example alcohol consumption has been coded as 1 for non-drinkers, 2 for drinkers, and named EXP. In GLIM terminology this variable is a FACTOR at two levels, that is a categorical variable with two categories. There is another possible type of explanatory variable recognised by GLIM, a VARIATE, which represents values of a continuous variable or an ordered categorical variable. The difference between these two types of explanatory variable arises when they are fitted in a model. GLIM fits constants to the values corresponding to levels of a FACTOR other than the first, while for a VARIATE a constant times the value of the variate is fitted, as in a linear regression model. A factor with only two levels does not require to be defined as such, since there is only one degree of freedom for the comparison of levels in any case. In general, factors may have several levels, m say, and these should be coded as the integers 1 to m inclusive.

The response, which I have named CASES, corresponds to the number of deaths in the appropriate exposure categories, and N is the total number of individuals in each exposure category. Only the first four characters of a name are recognised by the program. The $REAd directive precedes the data, and the program then reads the values to be placed in EXP(1), CASES(1), N(1) followed by those for EXP(2), CASES(2) and N(2).

Now the response variable and the type of regression model to be fitted are defined.

```
$YVAR CASES
$ERROR B N
```

These two directives inform the program that values of the response variable are held in the array CASES, and that the model to be fitted has a binomial error structure (B), with denominators corresponding to values of the variable in array N. The program automatically selects the appropriate default link function after the error structure has been defined, and

recognises that it is to fit a logistic linear regression to the proportions CASES/N. The null fit, that is fitting a common proportion to the two exposure categories is produced by the directive

```
$FIT
```

In order to obtain the output from the fit it is necessary to issue a new directive, which may in practice be null, as in

```
$FIT $
```

This will produce the following output from the program:

```
 scaled deviance = 2.1176 at cycle  3
             d.f. = 1
```

In order to examine the estimate of the combined rate, the directive $DISplay E is used. This directive takes a number of possible parameters, some of which will be described later. The parameter E requests estimates and standard errors, and produces

```
          estimate         s.e.      parameter
     1      -2.450       0.09805     1
```

The estimate of the common probability, $\hat{r}$ may be obtained by back-transforming from the logit as

$$\hat{r} = \frac{e^a}{1+e^a} = 0.079$$

where a is the estimate of α identified as corresponding to the parameter "1" in the output.

In this case the value of the deviance indicates that there is no significant difference in risk between the two exposure categories, as one would expect. It is for example only that we now fit a term in the model to take account of exposure to the risk factor.

```
$FIT EXP $

 scaled deviance = 0.00000000 at cycle  4
             d.f. = 0
```

The output shows a deviance of zero, but it is of interest to obtain the value of the parameter corresponding to exposure.

```
$DISPLAY E $

          estimate         s.e.      parameter
     1      -1.967        0.3376     1
     2      -0.2952       0.2010     EXP
```

By taking the exponential we obtain an estimate of the odds ratio due to drinking

$$\hat{\psi} = 0.744$$

1.8 Interval estimation

We may construct a confidence interval for the odds ratio by multiplying the standard error of the parameter estimate by the appropriate percentage point of the normal distribution and taking this distance either side of the estimate as the lower and upper limits. Back-transforming as above yields corresponding upper and lower limits for the odds ratio at the required level of confidence.

$(100-\alpha)\%$ confidence limits for $\log\psi$ are then given by

$$\log\hat{\psi} \pm z_{\alpha/2}.\ \mathrm{se}(\log\hat{\psi}) \qquad (1.8.1)$$

where $z_{\alpha/2}$ is the (a/2)% point of N(0,1).

Several different methods are available for constructing confidence limits appropriate to the comparison of two proportions, but since the present work is concerned with analysis of epidemiological studies by generalized linear modelling attention will be restricted to the above method, and to another simple method due to Miettinen (1976).

Miettinen's method involves only the estimate of $\log\psi$ and the value of the χ^2 statistic appropriate to the test of significance of exposure, χ_o^2. Arguing that the quantities

$$\frac{(\log\hat{\psi})^2}{\mathrm{var}(\hat{\psi})} \text{ and } \chi_o^2$$

provide equivalent information for the test of significance he suggests that $\mathrm{var}(\hat{\psi})$ be estimated by

$$\frac{(\log\hat{\psi})^2}{\chi_o^2} \qquad (1.8.2)$$

which gives as the interval for $\log\psi$

$$\log\hat{\psi} \pm z_{\alpha/2}.\ \frac{\log\hat{\psi}}{\chi_o} \qquad (1.8.3)$$

Back transformation of these limits gives

$$\hat{\psi}^{(1\pm z_{\alpha/2}/\chi_o)} \qquad (1.8.4)$$

These *test based limits* are particularly easy to apply in the 2x2 table, and do not involve any difficult calculation. Even in more complex analyses where variances might be difficult to estimate test-based limits are readily available if a suitable test has been performed.

Both the test based and standard error based limits are likely to be too narrow (Gart & Thomas, 1972; Gart, 1979), the former particularly so when calculated from small samples, and the latter when ψ is far from unity. Miettinen suggests using one of the test-based limits (upper or lower) if both the test based and standard error based limits are closer to the null value (ψ=1) than to the point estimate $\hat{\psi}$. In general however standard error based limits are more commonly used, and involve little extra computation given that a GLIM analysis is available.

Difficulties are posed by the presence of zero values in the cells of the table. In particular if no cases are observed both among those exposed to the risk factor and those unexposed it might be concluded that no information is provided by the study. If the total numbers in each exposure category are not equal then this conclusion is incorrect. For example 0 out of 10 is consistent with larger values of the unknown underlying probability than 0 out of 1000. Classical binomial analysis will provide interval estimates of the rate difference if desired. Needless to say the above data will not lead to a significant difference between the rates, but the idea that analysis of a study ends with calculation of a test statistic and a p-value is now recognised as inappropriate.

1.9 Extension to case-control studies

The data of GLIM example 1 were obtained from a longitudinal study of a sample from a specified population of individuals, and thus calculation of rates was possible. There should be no difficulty then in appreciating the validity of applying the logistic-linear model to the data. With the case-control study matters are not so clear-cut. The difficulty here is, as before, that the proportion of cases in a particular exposure category is not an estimate of the rate in that category. However the argument of section 1.4 can be extended to cover the case-control situation.

In the simple case of one exposure variable we may express the conditional probability of illness given exposure as

$$P(c/e) = \frac{P(e/c)P(c)}{P(e)} \qquad (1.9.1)$$

where c denotes the event "illness" and e exposure. This relation implies that the probabilities which are estimable from the case-control study, namely the conditional probabilities of exposure, given illness (for the cases), and absence of illness (for the controls) are related to the probabilities of illness given exposure or absence of exposure by multiplication by the appropriate constants. It is easy to see that these constants cancel in the odds-ratio as before. In order for this argument to be valid it is essential that the probability that an individual, whether a case or control, is selected is independent of exposure status, and that the selection process is independent for different individuals.

Although we have not yet considered the multivariate case in detail it is convenient to express the argument in terms of a vector of exposure variables, **x**. From a theoretical point of view a difficulty is posed by the undeniable fact that in a case-control study it is the exposure status of individuals, **x** which is a random variable, not their case status, which is controlled by the selection procedure. Denoting case or control status by y, and exposure status by **x** we can write

$$P(\mathbf{x}/y)=\frac{P(y/\mathbf{x})P(\mathbf{x})}{P(y)}; \ y=0,1 \qquad (1.9.2)$$

where P(y) corresponds to the probability of case (1) or non-case (0) status. The case-control study provides information about P(**x**/y), which is P(y/**x**), multiplied for cases by the overall probability of exposure divided by the probability of illness, and for controls by the probability of exposure divided by the probability of not having the illness. The logistic-linear model for P(y/**x**) appears therefore to require modification for application to case-control studies. However if the marginal distribution of exposure does not contain information about the parameters of the model for P(y/**x**) then it is possible to estimate P(**x**) and the coefficients of the model jointly. This approach leads to the same parameter estimates as applying the logistic linear model directly to the data.

An appropriate alternative is to use conditional likelihood methods to eliminate the term in P(**x**). This approach is based upon application of the logistic linear model conditional on the $\mathbf{x}_i$ values observed in the case and control groups in terms of their randomisation distribution. This will become clearer in the section on matched studies. The method leads to computational difficulties for large sample sizes, and asymptotically the resulting estimates will be close to those obtained by the joint likelihood method. A fuller theoretical treatment of these issues is provided by Farewell (1979).

GLIM example 2

The data of table 3 may therefore be analysed by exactly the same method as for table 2. It should be obvious from symmetry of the odds ratio implied in section 1.4 that the same results will be obtained whether one analyses the proportion of cases in each exposure category, or the proportions exposed in each status category, but for completeness both analyses will be given.

```
$UNITS 2
$DATA STATUS EXP N
$READ
2 298 767
1 87 323
$YVAR EXP
$ERROR B N
$FIT $
$FIT STATUS$
$DISPLAY E$
```

produces the following output

```
scaled deviance = 6.0723 at cycle  3
           d.f. = 1

scaled deviance = 0.0000000 at cycle  3
           d.f. = 0

         estimate         s.e.     parameter
    1      -1.349       0.2608     1
    2      0.3511       0.1444     STAT
```

The analysis of the proportion of cases in each exposure category is performed similarly.

```
$UNITS 2
$DATA EXP CASES N
$READ
2 298 385
1 569 805
$YVAR CASES
$ERROR B N
$FIT $
$FIT EXP$
$DISPLAY E$
```

```
 scaled deviance = 6.0723 at cycle  3
            d.f. = 1

 scaled deviance = 0.000000 at cycle  4
            d.f. = 0

          estimate         s.e.     parameter
```

```
1      0.5289      0.1971      1
2      0.3511      0.1444      EXP
```

Note the agreement between the two analyses, producing the same deviance statistics and estimated odds ratio.

II. MULTIVARIATE STUDIES

2. Multivariate studies.

There is little difficulty in analysing an epidemiological study involving only one dichotomous exposure variable and a similarly binary response. In practice studies usually include independent variables of various types, including categorical, ordered categorical and continuous variables. Some studies may also involve multiple response categories, but for simplicity we shall continue to consider dichotomous dependent variables only.

One reason for being interested in variables other than the exposure which is at the centre of an epidemiological enquiry is the possibility that other variables and their association with both exposure and the outcome variable may be influencing the relationship between exposure and illness.

2.1 Confounding

Consider an epidemiological investigation to assess the relative risk rr of developing an illness due to exposure to E. An extraneous variable X could obscure or exaggerate the relation between illness and exposure if it is associated with E and also with the probability of developing the illness. There is an interesting result concerning the relative risk of illness due to X if X is entirely responsible for an observed association between E and illness. The relative risk due to exposure to X must be at least as strong as that apparently due to exposure to E, and in addition X must be rr times more common among those exposed to E as in those unexposed. Taken together these two conditions are often considered to be good reasons for using relative risk as the measure of choice for relative disease incidence, large values of relative risk being unlikely to be due entirely to an unidentified confounding variable.

In the case of a single dichotomous possibly confounding variable we may represent the results of the study in a 2^3 table, and a suitable example of this is provided by the smoking and cataract analysis considered earlier, where we now tabulate by sex in addition to smoking.

Table 5: Smoking and cataract in Edinburgh by sex

		Cataract Cases	Population Controls	Totals
MALES				
	Smoker	184	33	217
	Non-smoker	135	64	199
	Total	319	97	416
FEMALES		Cataract Cases	Population Controls	
	Smoker	114	54	168
	Non-smoker	434	172	606
	Total	548	226	774

There are two questions of interest relating to sex as a confounding variable here. Firstly, does controlling for sex explain the association between cataract and smoking, that is, does the association observed in our earlier analysis arise simply because of differential rates of cataract and smoking between the sexes? If the answer to this question is in the negative then a further question is whether the association between smoking and cataract is the same for the two sexes.

It is not possible in this particular case to assess whether sex is a risk factor for cataract, inasmuch as defects in sampling the control population may have led to the difference in proportions of sexes between the case and control groups.

Again, although these data are from an actual study the analysis is presented here as an example of the methodology, not as a definitive solution to the problem of association between smoking and cataract. In fact this example will be considered in more detail in the next section.

We first specify what the above questions mean in terms of logistic-linear modelling. Simple inspection of table 5 shows that there are differing proportions of males and females between cases and controls. 37% of the cataract patients are males, whereas only 30% of the controls are. It should perhaps be mentioned that the cases for this study were persons who had cataractous lenses removed by surgery in Edinburgh, and tended to be rather elderly, and the controls were obtained from persons of similar age, hence the relatively low proportion of males in the groups. Sex may be associated with smoking habits, and also with the development of cataract, and therefore could be a confounding variable.

Examining the odds ratios in each of the two sexes from table 5 we see that for men $\hat{\psi}$ is estimated as 2.64, and for women 0.84. This would suggest that although sex may not be a confounder, it might be an effect modifier. We may nevertheless calculate a single estimate ψ for both sexes combined. This supposes that the true odds ratios in the two sexes are the same, and differ in the data only because of sampling variation.

For a series of 2x2 contingency tables, each classifying the cases and controls as in table 1, and each corresponding to a single level of a confounding variable with h levels, the Mantel-Haenszel estimate of the odds ratio is defined by

$$\hat{\psi}_{mh} = \frac{\sum_{i=1}^{h} a_i d_i / N_i}{\sum_{i=1}^{h} b_i c_i / N_i} \qquad (2.1.1)$$

In this formula a_i, b_i, c_i and d_i refer to the frequencies in the appropriate cells of the table corresponding to level i of the confounding variable. This estimator is not affected by the presence of zero values in some cells, and is consistent as the number of tables increases, even with small numbers in each table. Unfortunately this is not the case with estimators derived from logistic linear models for large numbers of sparse tables, but provided this drawback is recognised the logistic linear model is more convenient for many purposes.

So, for the data in table 5, calculation gives

$$\hat{\psi}_{mh} = 1.31 \qquad (2.1.2)$$

GLIM example 3

One suitable (and simple) formulation of a logistic model for these data is to take as response the numbers of cases in each exposure category by sex, with denominator given by the total number of cases in the sex by exposure category, and include terms in the model representing sex and exposure. As there are 4 proportions to be modelled and two parameters to be fitted, namely sex and smoking, there will be one degree of freedom after fitting, and the residual deviance associated with this is a suitable test statistic for effect modification.

The GLIM control language to perform the analysis is

```
$UNITS 4
$DATA EXP SEX CASES N
$READ
2 1 184 217
1 1 135 199
2 2 114 168
1 2 434 606
$FAC EXP 2 SEX 2
$YVAR CASES
$ERROR B N
$FIT EXP+SEX$
$DISPLAY E$
```

The output from this analysis shows a residual deviance of 14.20 on 1 df, which is evidence of lack of fit of the model, strongly suggesting effect modification by sex. Residuals from the fitted model may be inspected by an additional option in the $DISPLAY directive as

```
$DISPLAY E R$
```

producing the following output

```
        estimate          s.e.      parameter
  1       0.9870        0.3634      1
  2       0.2825        0.1513      EXP
  3      -0.2211        0.1471      SEX
  scale parameter taken as   1.000

 unit  observed     out of     fitted    residual
   1        184        217      171.6       2.062
   2        135        199      147.4      -1.998
   3        114        168      126.4      -2.207
   4        434        606      421.6       1.091
```

The effect estimates above (0.2825 for exposure and -0.2211 for sex) are of little interest, since sex appears to modify the effect of smoking, but it is worth remarking that transforming the smoking effect from log odds ratio to odds ratio gives an estimate

$$\hat{\psi}=1.32$$

very close to the Mantel-Haenszel estimate.

The residuals from the model serve as another indication that sex is an effect modifier. The observed number of smoking cases among the males is considerably greater than the fitted number, the converse being true for females. The residuals themselves are standardised by the estimated binomial standard deviation of the observations in each category, as

$$\frac{\text{observed}_i\text{-fitted}_i}{\sqrt{\text{fitted}_i\,\{1-\frac{\text{fitted}_i}{\text{total}_i}\}}} \qquad (2.1.3)$$

where each of the quantities in (2.1.3) refers to the individual data point corresponding to category i.

The use of residuals in GLIM is much the same as in standard regression analysis. Examination of residuals may suggest ways to improve model adequacy, identify outliers and so on. Just as in regression analysis there are several different types of residuals which may be calculated using specially written GLIM macros. Many standard regression diagnostic procedures may be modified to be applied to the logistic-linear model (Williams, 1987).

Notice that the $FIT directive takes as arguments the names of the independent variables to be entered in the model. Subsequent fits may be specified by updating the directive, for example, if we now wished to fit exposure alone this could be done by issuing the directive

```
$FIT -SEX$
```

while to fit the interaction of sex and exposure we could update the (initial) model formula by

```
$FIT +EXP.SEX$
```

2.2 Confounding variables at several levels

Rather than immediately estimating the differential effect of smoking as a risk factor between the sexes it may be more useful to expand the data to take account of another factor. The age structure differed between cases and controls, and although it seems likely that age is a risk factor for the development of cataract, again it is being treated here as a potential confounder, because difficulties in control sampling may have led to the different distributions of cases and controls by age. To distinguish factors of this type from the exposure factors of primary interest they will be referred to as stratifying variables or stratifiers.

Table 6: Smoking and cataract by age and sex

		Males		Females	
Age	Smoking	Cases	Controls	Cases	Controls
50-59	No	11	6	17	23
50-59	Yes	42	1	18	8
60-69	No	38	26	72	59
60-69	Yes	59	11	41	23
70-79	No	60	28	191	70

70-79	Yes	60	20	47	22
80-89	No	26	4	154	20
80-89	Yes	23	1	8	1

For convenience and to illustrate another GLIM feature the table summarising smoking, age and sex is presented as numbers of case and controls classified by smoking and the stratifiers rather than as a series of 2x2 contingency tables with marginal totals. It will be noticed that, as mentioned previously, the cases and controls are all rather elderly. Although some cells contain very low numbers there are no zeroes in the table, so log odds ratios for all the subtables by stratifiers are estimable, and may therefore be combined without difficulty, should it prove reasonable to do so.

Before performing the GLIM analysis it is necessary to give some guidance regarding model selection. Clearly one can include terms involving age (3 degrees of freedom), sex (1 df), smoking (1df), and the interactions of these, age by sex (3 df), age by smoking (3df) sex by smoking (1df) and age by sex by smoking (3 df). Of these terms the ones of direct interest are those which involve smoking, and the interaction age by sex is of no direct relevance to the investigation. However, since what we are trying to do is to take account of possible confounding due to age and sex it seems reasonable to include all terms in the stratifying variables, including their interaction in any model.

Now, there are 16 observed proportions, and thus 15 df in the data. The degrees of freedom corresponding to the terms listed above also add to 15, so fitting all terms leads to a residual deviance value of zero. As in the last example then we may fit all terms *except the highest order interaction*, and see if this leads to a satisfactory model, as judged by comparing the residual deviance with χ^2 on 3 df. If the value of the residual deviance is larger than the selected percentage point, then this is evidence of a rather complicated system of effect modification of smoking, in such a way that the odds ratio due to smoking differs between age by sex classes in a way that cannot be expressed as a product of separate marginal estimates for age and sex. If this should be the case then modelling has not simplified the interpretation of the data- not necessarily through any defect of the model, but because the data indicate that the underlying population odds-ratios in different age-sex groups do not have a structure that permits simple description.

On the other hand, if the highest order interaction is not required, terms involving the exposure factor may be deleted, one at a time, starting with the highest order terms, and the resulting increases in deviance assessed against the χ^2 distribution with the appropriate degrees of freedom. This method of model selection is a constrained form of backward deletion, a stepwise method described in some detail for the linear regression case in Daniel & Wood (1971). It is constrained in that we do not assess the effect of deleting terms involving stratifiers only. The process of deleting terms from the model continues until a term is deleted which increases the residual deviance too greatly. If this term is an interaction of the exposure factor with a stratifier then it is of interest to assess all the other interactions of exposure with stratifiers of the same order. In some cases with exposure factors which are ordered categories it will also be useful to test whether a single regression coefficient may be applied to the factor rather than fitting separate constants to each level. This will become clearer as we go through the examples in the course.

GLIM example 4

The GLIM control language required to perform the analysis of the data in table 6 along the lines outlined above is as follows.

```
$UNITS16
```

```
$DATA AGE EXP SEX CASES CONTROLS
$READ
1 1 1 11 6
1 1 2 17 23
1 2 1 42 1
1 2 2 18 8
2 1 1 38 26
2 1 2 72 59
2 2 1 59 11
2 2 2 41 23
3 1 1 60 28
3 1 2 191 70
3 2 1 60 20
3 2 2 47 22
4 1 1 26 4
4 1 2 154 20
4 2 1 23 1
4 2 2 8 1
```

The above lines define and read in the data. It remains to declare the factor structure, compute the denominator for the cases, and set the error structure and response variate.

```
$FACTOR AGE 4 SEX 2 EXP 2
```

It is not strictly necessary to declare EXP and SEX as factors at two levels as explained earlier, but AGE must be specified as having four levels.

```
$CALCULATE N=CASES+CONTROLS
```

This illustrates the use of the $CALculate directive, which as its name implies is used to perform calculations, transformations and logical comparisons on the data. Notice that the vector N which will hold the totals of CASES and CONTROLS for each cell of the classification has not been previously defined. The directive creates a vector called N without requiring any further specification.

```
$YVAR CASES
$ERROR B N
```

In the specification of the logistic-linear model interactions are written as terms in the products of explanatory factors, and in a GLIM fit a single interaction of two factors may be represented as for example

```
AGE.SEX
```

An interaction of three factors may be written as

```
AGE.SEX.EXP
```

and using this notation the model to be fitted to the present set of data can be expressed as

```
AGE+SEX+EXP+AGE.SEX+AGE.EXP+SEX.EXP
```

An alternative notation which is shorter but means the same thing is provided in GLIM using * instead of . between factor names. It is quite easy to understand, in that

```
AGE*SEX ≡ AGE+SEX+AGE.SEX
```

Since in the present example we wish to include the three terms on the right hand side in all models to be fitted we may use this notation in the $FIT directive. Brackets may also be used to economise in the length of specification of a model. These follow the normal rules of brackets of multiplication. Thus

```
EXP.(AGE+SEX) ≡EXP.AGE+EXP.SEX
```

The appropriate directive to fit all terms except the highest order interaction to the present data is therefore

```
$FIT AGE*SEX+EXP+EXP.(AGE+SEX) $
```

Of course, the directive may also be used with the individual terms listed rather than the short cut notation, and the result will be the same.

The residual deviance is 1.42 on 3 df, which indicates that the three-factor interaction EXP.SEX.AGE is not required in the model. We now examine each of the two-factor interactions involving exposure in turn, making use of the model formula updating facility in GLIM by which terms may be added to or deleted from the current model without the necessity of giving a full model specification.

```
$FIT -EXP.SEX$

scaled deviance = 8.6 (change = +7.18)
                   df = 4    (change = +1)
```

The increase in deviance resulting from the exclusion of the exposure by sex interaction is much greater than the 1% point of χ^2 on 1 df, and we conclude that the exposure by sex interaction must be retained in the model.

To assess the importance of the exposure by age interaction we examine the effect of deleting this from the initial model. One way to do this quickly is to add back the term EXP.SEX and remove EXP.AGE

```
$FIT +EXP.SEX-EXP.AGE$
```

The resulting deviance and df are to be compared with the original values 1.42 and 3, not the last values, so the change values given in brackets in the output are ignored.

```
scaled deviance = 14.79
                   df =   6
```

Thus, deleting the exposure by age interaction increases the deviance by 13.37 on 3 degrees of freedom, and we conclude that this interaction is also required in the model.

The effect estimates and standard errors are as follows.

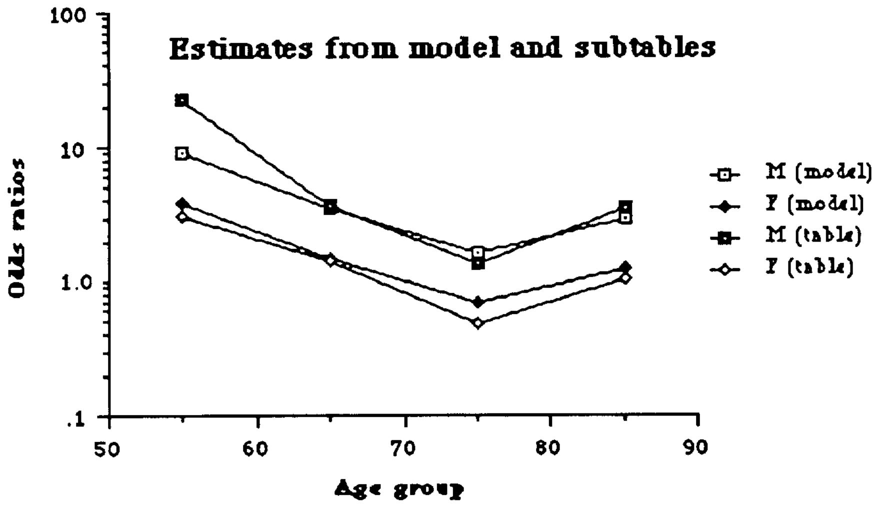
```
estimate        s.e.        parameter
  0.8479       0.4657       1
 -0.4514       0.5098       AGE(2)
 -0.1536       0.4990       AGE(3)
   1.067       0.6897       AGE(4)
  -1.243       0.5149       SEX(2)
   2.225       0.5236       EXP(2)
   1.038       0.5614       AGE(2).SEX(2)
   1.578       0.5475       AGE(3).SEX(2)
   1.361       0.7501       AGE(4).SEX(2)
 -0.9699       0.5319       AGE(2).EXP(2)
  -1.731       0.5207       AGE(3).EXP(2)
  -1.149       0.9304       AGE(4).EXP(2)
 -0.8505       0.3187       SEX(2).EXP(2)
```

These require some explanation, and care in interpretation. First of all it should be recalled that age and sex are considered here as confounders and effect modifiers, not as risk factors in their own right. The apparent association between these factors and cataract may be due to deficiencies in control sampling and thus we do not attempt to interpret estimates for age, sex or age by sex.

On the other hand we must take account of all estimates involving the exposure variable. The estimate of the log-odds ratio for exposure is 2.225, which is the estimate of the effect of exposure at the lowest levels of all other factors in the model. In the present case therefore it corresponds to an estimated log odds-ratio for exposure among men aged 50-59. The estimated log odds-ratios for men in the other age groups are obtained by adding the appropriate estimates of components of the exposure by age interaction. Thus the estimated log odds-ratio for men aged 70-79 due to exposure is 2.225-1.731 which equals 0.494. Figure 1 shows the estimated odds ratios from the model and the odds ratios calculated from the appropriate subtables for each age group and sex. Notice that the estimated log odds-ratios for women in the various age groups are obtained by adding the sex(2).exp(2) estimate to the corresponding values for males.

The estimates from the model and the calculated odds ratios from the subtables are in quite close agreement except for the youngest males. The reason why this disagreement does not lead to significance of the three factor interaction is that there are only 7 control subjects for this category. It is interesting that for women

Figure 1: Cataract and smoking by age and sex



aged 70-80 exposure to smoking appears to reduce the risk of cataract. However the standard errors in the output show that the actual value of log odds ratio due to exposure for this group is not significantly different from zero. It may be concluded that men are placed at more risk by smoking than are women, and that the risk related to smoking varies by age in a similar fashion for each sex.

2.3 Independent variables- VARIATEs of FACTORs?

As mentioned in section 1.7 independent variables in the regression model may declared as either FACTORs or VARIATEs. For a factor x at m levels, $x_{1,} x_2, \ldots, x_m$ GLIM fits m-1 constants to the levels 2,3,... m of the factor. In the case of a variate a single regression coéfficient β is fitted. The models for these are as follows, with y representing the logit of the appropriate proportion:

$$E[y/x_k] = \alpha+\beta_k \; ; \; k=2,3,\ldots m \qquad (2.3.1)$$

$$E[y/x] = \alpha+\beta x \qquad (2.3.2)$$

These are simply extended to vectors of dependent variables, which may include both factors and variates. However in modelling response it is not possible to use the interaction symbols to join two variates. To fit an interaction of two variates, x_i and x_j we must define a new variate, using the $CALculate directive, having values given by the products $x_i x_j$. Terms involving the interaction of a single variate with one or more factors may be specified in the usual notation. The test of this type of interaction in a logistic-linear model is analogous to a test for parallelism of regressions in standard regression analysis. Polynomial terms may be fitted by creating new variates having the values of the appropriate powers of the variates concerned.

As anexample of these approaches consider the data in table 7 which are taken from a study of liver cirrhosis and alcohol consumption in Ille et Vilaine (Pequignot et al, 1978).

Table 7: Ascitic cirrhoses and controls by daily alcohol consumption

Daily consumption (in gms 100% alc)	Ascitic Cirrhoses	Controls
0-20	3	185
21-40	10	212
41-60	15	165
61-80	24	108
81-100	30	58
101-120	23	31
121-140	25	13
141-160	24	5
161+	30	1

Table 7 represents the distributions by alcohol consumption category of 184 male sufferers from alcoholic cirrhosis and 778 male controls obtained from the general population. The 9 consumption categories may be considered as 9 levels of a single factor, or we may associate with each category a value of consumption. For example, the first category may be taken to correspond to 10 gms/day, the second 30 gms/day, and so on. There is clearly some arbitrariness in this procedure, which is particularly evident when considering the last category, which has no upper limit. Nevertheless, it is a useful exercise to investigate the feasibility of replacing the 8 degrees of freedom factor with a single degree of freedom variate. The success or failure of the simpler (variate) formulation may be assessed by comparison of the deviance difference between the models with the appropriate percentage point of χ^2 on 7df. However in the present case fitting alcohol consumption as a factor at 9 levels will result in a residual deviance of zero, with zero degrees of freedom, and we therefore appropriate to perform a single analysis with a suitably defined variate.

GLIM example 5

The GLIM control language required is

```
$UNITS 9
$DATA ALC CASES CONTS
$READ
1 3 185
2 10 212
3 15 165
4 24 108
5 30 58
6 23 31
7 25 13
8 24 5
9 30 1
$CAL TOT=CASES+CONTS
$YVAR CASES
$ERROR B TOT
$CAL VALC=20*ALC-10
```

This assigns values 10,30,50,....,170 to the elements of a new vector VALC according to the corresponding value of ALC. The null fit may be obtained by

```
$FIT $

scaled deviance = 327.34
                       df = 8
```

The value of the deviance shows a very significant relationship between alcohol consumption and cirrhosis, but we now attempt to explain this in terms of VALC, that is consumption level considered as a variate.

```
$FIT VALC$

scaled deviance = 3.4372
                       df = 7
```

Thus almost all the variation between the alcohol consumption categories can be explained by logistic-linear regression on VALC.
The effect estimates are obtained as

```
    estimate       s.e.     parameter
      -4.327     0.2525             1
     0.03934   0.002796          VALC
```

Different, but sensible values chosen to represent the maximal consumption category produce roughly similar values of the deviance and the estimated regression coefficient. The interpretation of the coefficient of VALC in this analysis is that each gram of alcohol consumed increases the estimated log odds ratio of cirrhosis by 0.03934. To put this in a more concrete way, a person drinking an extra 20 grams of absolute alcohol per day multiplies his odds ratio of cirrhosis by approximately 2.2, and for low values of initial risk this is approximately equivalent to risk multiplication by the same factor.

A further example may be considered using a more detailed tabulation of the data in table 2, presented in table 8.

Table 8: 10-year deaths of English civil servants by age and alcohol consumption

	Alcohol (gm/day)			
Age (yr)	0	0.1-9	9.1-34	>34
40-49	6/206	9/174	7/144	3/60
50-59	26/206	8/172	13/177	12/105
60-64	13/65	7/44	4/46	5/23

The entries in the table represent the numbers of deaths in each particular age and consumption category, and the total numbers of men in the category found in the survey. Once again the assignment of values to each consumption category is somewhat arbitrary. In the reported analysis (Marmot et al, 1981) the values used are not given explicitly, but from a figure in the text it appears that equal intervals between the values for the four categories were employed. This does not make a great deal of sense. The values used in the present analysis were 0, 5, 23 and 45.

GLIM example 6

To analyse the above data along the lines indicated, declare 12 units, factors AGE (at 3 levels) and ALC (at 4). Read in the data, declare a binomial error structure and fit the model excluding the two-factor interaction.

```
$FIT AGE+ALC$

scaled deviance = 8.70
                     df = 6
```

This indicates that the AGE.ALC interaction is not required in the model. We may fit the models AGE alone and ALC alone, and doing so will indicate that ALC can in fact be excluded, χ^2 = 5.2 on 3 df.

```
$FIT AGE$

scaled deviance = 13.96
                     df = 9
```

Rather than leave matters there, we may consider fitting alcohol consumption as a variate on one degree of freedom, and fitting polynomial terms in consumption. Following the line of the original authors consumption categories were identified with variate values but in the present analysis these were assigned to correspond more closely to the actual categories.

```
$CAL VALC=0
$CAL VALC=%IF(ALC==2,5,VALC)
$CAL VALC=%IF(ALC==3,23,VALC)
$CAL VALC=%IF(ALC==4,50,VALC)
```

Notice the rather unusual structure of these assignments. The syntax of conditional assignment in GLIM involves a variable to be assigned, an operator, and 3 arguments in brackets. The first argument is usually a logical variable, or expression, taking values 1 (true) and 0 (false). The second argument is the value taken by the assigned variable when the first argument is 1, the third being the assignment when the first parameter is 0. Thus the last statement above means

When ALC has the value 4, assign the value 50 to VALC, otherwise do not change the value of VALC.

We may now proceed to fit AGE+VALC, with the following result

```
scaled deviance = 13.84
                     df = 8
```

and again it is clear that VALC is not required in the model. However we may fit a quadratic term in VALC, V2

```
$CAL V2=VALC*VALC
$FIT AGE+VALC+V2
```

```
scaled deviance = 9.981
                      df = 7
```

Now we have a reduction in deviance due to linear and quadratic terms in VALC of approximately 4 on 2 df. Again, we should conclude that these terms are not required in the model. But things are not quite so simple. Addition of the quadratic term leads to a deviance reduction of 3.85 on 1 df which is greater than the 5% point of χ_1^2 indicating that the quadratic term should be included in the model. Examining the data, and the estimates from the GLIM analysis shows that the odds ratios for the highest category compared to the lowest is approximately 1, and the two intermediate categories have odds ratios of less than 1, in fact about 0.65 relative to the zero category. So, if the quadratic term is included the effect is to suggest only that intermediate consumption is beneficial. The danger of this result is that extrapolation of the regression model to values of consumption higher than that assumed to correspond to the highest observed category could lead to the unjustified assertion of very high odds ratios for heavy consumers of alcohol.

2.4 Analysis of ungrouped data

The previous examples involving variates proceeded on the basis of grouped data, and the assignment of a group mid-point value of the variate to each group. In practice it is often the case that the individual values of the variate are available for each unit, and in these circumstances it makes sense to use them. The data may then be entered into GLIM without grouping, individual by individual. The response variable will be case status of the individual, taking values 0 (control) or 1(case), and the appropriate error structure is binomial, with a denominator of 1. If there are a large number of factors in the data it may sometimes be the case that grouping on these may lead to many small strata, and information may not be available in the data concerning particular effects or interactions. There may also be many parameters to estimate in taking account of the stratifying variables, and if the number of these parameters is close to the number of observations effect estimates can be seriously biased.

Sparse data then can lead to problems of interpretation, and the result of an analysis of ungrouped data should never be assessed purely on the deviance differences. Effect estimates and their standard errors are in any case of more interest than significance levels, but it is also important to examine the data to ensure that categories contain reasonable numbers of observations. If not, it is necessary either to use a conditional logistic regression approach, which is not easily implemented, or to combine classes of the stratifying variables to increase the numbers of observations in each stratum and reduce the number of nuisance parameters (effect estimates for stratifiers).

GLIM example 7

An example of analysis of data of this type is provided by a study in Edinburgh of alcohol consumption and problems among brewers and company directors. Several variables were measured for each of the selected individuals, but we shall concern ourselves here with only three. The data in Appendix A concern non-abstinent company directors interviewed for the study, and include the weekly alcohol consumption of the respondents (in units approximately equal to 1cl of 100% alcohol), the number of days of the week on which respondents drank and the experience by respondents of symptoms of alcohol dependence.

The data for the 223 respondents may be entered from a computer file using the GLIM directive $DINPUT nn, where nn is a channel number which will be associated with the data file. The program prompts for a file name which should be entered in the manner appropriate to the computer system being used. The variable names used are AP for experience of alcohol dependence symptoms, WKU weekly alcohol consumption and DS, number of drinking days. The control language and analysis is as follows.

```
$UNITS 223
$DATA WKU DS AP
$DINPUT 11
B:DIRDRK.DAT
$CAL N=1
```

This sets the binomial denominator N (which will be declared in the $ERROR directive) equal to 1 for each individual.

```
$FAC DS 7
$YVAR AP
$ERROR B N
```

We now fit the full model with alcohol consumption as a variate, and number of drinking days as a factor at 7 levels.

```
$FIT DS+WKU+DS.WKU$
```

The scaled deviance output is of use only as a marker against which to assess the effect of deleting variables. It has no other interpretation in terms of the goodness of fit of the model. (Williams, 1987).

```
  scaled deviance = 126.55 at cycle  8
             d.f. = 209

 $FIT -DS.WKU$

  scaled deviance = 140.10 (change =  +13.55) at
cycle  4
             d.f. = 215    (change =   +6    )
```

Hence the interaction of drinking days and consumption is required in the model, and in these circumstances model selection need proceed no further. Adding back this interaction and displaying the effect estimates gives

```
$FIT +DS.WKU$

  scaled deviance = 126.55 (change =  -13.55) at
cycle  8
             d.f. = 209    (change =   -6    )
```

```
$DIS E$
          estimate          s.e.      parameter

    1      -5.220          1.770      1
    2      -8.165          13.79      DS(2)
    3       2.353          2.001      DS(3)
    4       2.457          2.202      DS(4)
    5       3.034          2.062      DS(5)
    6      0.9411          2.367      DS(6)
    7       1.842          2.028      DS(7)
    8      0.1769        0.09299      WKU
    9     0.08198         0.2863      DS(2).WKU
   10     -0.1748        0.09474      DS(3).WKU
   11     -0.1320        0.09736      DS(4).WKU
   12     -0.1746        0.09476      DS(5).WKU
   13     -0.1306        0.09558      DS(6).WKU
   14     -0.1540        0.09338      DS(7).WKU
```

Although it should be clear enough from the parameter estimates that the relationship of syptomatology and drinking days is not linear on the logistic scale, it is of interest to confirm this by fitting DS as a VARiate, and defining a new variate to represent the interaction between consumption and DS.

```
$VAR DS
$CAL XDS=DS*WKU$
$FIT WKU+DS+XDS$

  scaled deviance = 147.46 at cycle  4
             d.f. = 219
```

The increase in deviance is 20.91 on 10 degrees of freedom, hence the non-linearity of the relationship is confirmed.

It is not the intention of the present course to attempt to explain and interpret the substantive findings of the example analyses, but it is important to know how to interpret the effect estimates. These may be most easily represented on a graph which relates the log odds ratios of experiencing the symptoms to amount consumed, with separate lines for each of the seven drinking days categories. For those with only one drinking day, the slope of the line is 0.1769, while for those with 7 drinking days the slope is much less, 0.1769-0.1540, only 0.0229, but the intercept of this line is 1.842 above that for one drinking day. For convenience the intercept of the first line may be considered zero.

2.5 Analysis of matched pairs

An instance where conditional methods are easy to use in GLIM is the analysis of matched sets, and for simplicity this is illustrated by consideration of the matched pair case-control study. As the name suggests, the design of the study is that each case is individually matched with a control on a number of characteristics which are thought to be potential confounders. A case-control pair can therefore be thought of as a single stratum, where the strata are defined in terms of combinations of levels and values of the matching variables. The exposure status of case and control is recorded and a summary table can be drawn up as in section 1.5. For analytic purposes the data are best considered as a set of individual records, one for each case-control pair.

The theoretical basis assumes that the logistic-linear model is appropriate to the probability of being a case, that is for an individual with values $\mathbf{x}_i$ of the explanatory variables

$$\mathrm{P}[\text{case}/\mathbf{x}_i] = \frac{\exp(\alpha+\beta'\mathbf{x}_i)}{1+\exp(\alpha+\beta'\mathbf{x}_i)} \qquad (2.5.1)$$

To construct the conditional likelihood for a single case control pair we may write $\mathbf{x}_{i1}$ and $\mathbf{x}_{i0}$ as the vectors of explanatory variates for the case and control members of the pair respectively. The conditional likelihood of interest is then the probability that of the two members of the pair the $\mathbf{x}_{i1}$ corresponds to the case and $\mathbf{x}_{i0}$ to the control. From the model the log odds of being a case for each of these are given by

$$\exp(\alpha+\beta'\mathbf{x}_{i1}) \text{ and } \exp(\alpha+\beta'\mathbf{x}_{i0}) \text{ respectively.}$$

Hence the log odds-ratio within each pair for the observation is given by

$$\exp(\beta'(\mathbf{x}_{i1}-\mathbf{x}_{i0})) \qquad (2.5.2)$$

which may be fitted in the usual way, provided that the intercept (in GLIM model notation "1") is omitted from the models.

Alternatively, the probability of being a case for each member of the pair is given by

$$\frac{\exp(\alpha+\beta'\mathbf{x}_{i1})}{1+\exp(\alpha+\beta'\mathbf{x}_{i1})} \text{ and } \frac{\exp(\alpha+\beta'\mathbf{x}_{i0})}{1+\exp(\alpha+\beta'\mathbf{x}_{i0})} \qquad (2.5.3)$$

respectively. Thus the required probability is

$$\frac{\dfrac{\exp(\alpha+\beta'\mathbf{x}_{i1})}{1+\exp(\alpha+\beta'\mathbf{x}_{i1})}\;\dfrac{1}{1+\exp(\alpha+\beta'\mathbf{x}_{i0})}}{\dfrac{\exp(\alpha+\beta'\mathbf{x}_{i1})}{1+\exp(\alpha+\beta'\mathbf{x}_{i1})}\;\dfrac{1}{1+\exp(\alpha+\beta'\mathbf{x}_{i0})}+\dfrac{\exp(\alpha+\beta'\mathbf{x}_{i0})}{1+\exp(\alpha+\beta'\mathbf{x}_{i0})}\;\dfrac{1}{1+\exp(\alpha+\beta'\mathbf{x}_{i1})}} \qquad \ldots(2.5.4)$$

which can easily be seen to simplify to

$$\frac{1}{1+\exp(\beta'(\mathbf{x}_{i0}-\mathbf{x}_{i1}))} \qquad (2.5.5)$$

equivalent to the usual form of logistic model probability

$$\frac{\exp(\beta'(\mathbf{x}_{i1}-\mathbf{x}_{i0}))}{1+\exp(\beta'(\mathbf{x}_{i1}-\mathbf{x}_{i0}))} \qquad (2.5.6)$$

in which there is no intercept term and the explanatory variates are the within-pair differences (case - control) on the independent variates. If there are n pairs the likelihood is simply the product of the n expressions above for each pair. Notice that variates which are used in the matching process will be identically zero for all pairs, and thus their influence on risk cannot be estimated. We may assess whether they modify the effects of exposure by forming the appropriate products as variates or factors to be entered in the model. Factors pose a slight problem, in that if the factor levels are truly qualitative the within-pair differences between these are not suitable measures. For example with a four level factor a case control pair difference of 1 between levels of the factor could arise in three different ways, with different interpretations. This may be dealt with by creating dummy variables corresponding to combinations of factor levels in the pairs.

A GLIM analysis proceeds by entering the data as within-pair differences, declaring a YVAR equal to 1 in all cases, and fitting models each of which excludes the intercept, by specifying "-1" in the $FIT directive. The "null fit" for this approach is obtained by declaring a vector of zeroes and fitting this, without an intercept (equivalent to a probability of 0.5 for each pair).

GLIM example 8

A simple example of a matched study is provided by the following data based on a psychiatric study in Edinburgh of young women who had undergone appendectomy, but were found to have normal (non-inflamed) appendices. 39 such women were individually matched on the bases of age, social class and educational background with 39 women from a community sample. All subjects were interviewed, and their experience of life events, difficulties and level of social support were ascertained. The social support measurement was actually an ordered qualitative variate (a factor), but is coded for simplicity as zero when both case and control had the same level of social support, 1 when the case had more support than the control, and -1 when the control member of the pair had more support than the case. The analysis undertaken here does not attempt to assess the influence of interaction terms. However this can be done if required by forming appropriate interaction variables separately for the case and control members of each pair and taking the within-pair differences in the usual way.

The data are contained in Appendix B, and have been changed from those in the original study (Vassilas, M.Phil. Thesis, 1988, unpublished). For the purpose of the GLIM analysis the data have been entered as within-pair differences, case-control, and stored in the file PAIRDAT. Variable names used are EVENT, for differences in experience of life events, DIFF for differences in experience of difficulties, and SUPPORT for differences in level of social support available.

```
$UNITS 39
$DATA EVENT DIFF SUPPORT
$DINPUT 10 $
 File name? B:PAIRDAT
$CAL ZERO=0
$CAL R=1
$CAL N=1
$YVAR R
$ERROR B N
$FIT ZERO-1$
```

This null fit is of interest only as a reference point for evaluating the contributions of the variables to be fitted.

```
scaled deviance = 54.065 at cycle  2
            d.f. = 39
```

```
$FIT EVENT+DIFF+SUPPORT-1$
```

```
scaled deviance = 42.019 at cycle  4
           d.f. = 36
```

The deviance reduction of about 12.0 on 3 degrees of freedom indicates that one or more of the fitted variables is significantly associated with appendectomy. In order to simplify the process of fitting and reducing the number of terms in the model it is useful to examine the effect estimates.

```
$DISPLAY E$
```

```
        estimate          s.e.      parameter
  1        1.734        0.6397      EVEN
  2       0.2070        0.5941      DIFF
  3      -0.1579        0.5656      SUPP
```

It appears that difficulties and social support are not significantly associated with appendectomy, and although elimination of one of these could increase the significance of the other, in this case elimination of both shows that neither is required in the model.

```
$FIT EVENT-1$

scaled deviance = 42.178 at cycle  4
           d.f. = 38
```

Clearly there is no question of either of the individual χ^2 statistics for DIFF or SUPPORT being significant when together they only reduce the deviance by 0.157. For completeness it would be desirable to assess each on its own compared to the null fit, but for brevity this is left to the reader. The estimated log odds ratio corresponding to experience of a life event is obtained by

```
$DISPLAY E$

        estimate          s.e.      parameter
  1        1.791        0.6154      EVEN
```

which can be transformed to give an odds ratio estimate of 6.00, obtainable directly by applying the formula given by equation (1.5.1) to the data for event differences.

2.6 Multivariate modelling and joint action.

There has been a great deal of interest over the past fifteen years or so in the concepts of synergism and antagonism between causal agents in epidemiological research, and this has led to attempts to characterise certain modes of joint action as synergistic or antagonistic. For example Brown & Harris (1978) analysed data concerning the incidence of depression in women in Camberwell, London and claimed to have detected an interaction between the risk factors "experience of a severe life event" and "lack of intimacy" (in the sense of a confiding relationship with a partner).

Some of their data on this subject are represented in table 9 (taken from Everitt & Smith, 1979).

Table 9: Depression in women in Camberwell

	Lack of intimacy			
	Yes		No	
Severe eventYes	No	Yes	No	
Case	24	2	9	2
Non-case	52	60	79	191
Total	76	62	88	193
Rates	0.32	0.03	0.10	0.01

Since this was a community study the rates may be directly estimated by the proportions in each column. It can be seen that the relative risks due to experience of a severe event are approximately equal in each of the intimacy categories, and symmetrically, that the relative risks due to lack of intimacy are similar in each of the event categories. Thus, in terms of relative risk lack of intimacy is not a modifier of the "effect" of a severe event. However if we consider rate differences rather than ratios the rate difference due to experience of a severe event in those who lack intimacy is 0.29 compared with only 0.09 among those who have an intimate relationship. On an additive interpretation of effect, lack of intimacy *is* an effect modifier.

Brown & Harris noticed that among those not experiencing a severe event the χ^2 statistic for the relevant 2x2 table was not significant. From this they concluded that lack of intimacy was not itself a risk factor, and led to increased risk only in conjunction with a severe event. They accordingly identified lack of intimacy as a "vulnerability factor", whereas experience of a severe event was considered a "provoking agent".

Statistical interaction depends on the scale of measurement. In the present example a multiplicative measurement of effect as in a logistic-linear model fits the data adequately without the need for an interaction term, whereas an additive model would require such a term. The scale of measurement of effect and model for combination of effects cannot be entirely at choice, since an additive model of risk difference may lead to difficulties associated with all risks involved in the model being required to lie between 0 and 1 as they are probabilities. Similarly, a multiplicative model for relative risk could lead to absolute risks of greater than 1, whereas a multiplicative odds-ratio model such as the logistic-linear model will not lead to this inconsistency.

The application of statistical methods to data of this type cannot therefore be expected to yield definitive conclusions regarding synergism, antagonism or their absence. Rothman (1986) proposes a model for interaction which takes as axiomatic the measurement of effect in terms of rate difference. Walter & Holford (1978) present models of causation and joint action which give rise to additive, multiplicative and other models of joint action, showing in effect that there is no inherently "natural" way in which causal agents act together.

Causal models giving rise to various forms of joint action can be elaborated, and to some extent tested. However it is not always possible to test these with any degree of power on the basis of a single epidemiological study. In the example considered it seems that the main qualification of lack of intimacy as a vulnerability factor is that it does not increase the risk of depression except in conjunction with the provoking agent. One might therefore attempt to assemble a large sample of individuals free of the provoking agent, in order to perform a sensitive test of the potential effect of the vulnerability factor (a negative result being essential to the hypothesis). This approach will not work if there is misclassification of individuals as not exposed to a provoking agent when they are so exposed. It will also be ineffective if there exist other provoking agents which are unknown and which are potentiated by lack of intimacy. In both these cases there will be a higher rate of depression among those lacking intimacy because of the presence in the study of individuals exposed to an undetected provoking agent.

2.7 Analysis of vital rates

The comparison of mortality and morbidity rates between countries at a single point in time or within a country over several time periods forms an important part of descriptive epidemiology. Such comparisons may lead to hypotheses concerning possible risk factors which may subsequently be tested by more direct investigation, quite apart from the importance of monitoring rates over time in a single country for administrative reasons.

Mortality rates may be considered as realisations of binomial random variables and logistic-linear modelling may be applied using the approach of Chapter 2, but it is sometimes useful to model vital rates using a Poisson error structure with the logarithm of total population at risk as an offset, and a logarithmic link function. The results of this approach are exactly similar to the more obviously appropriate binomial model.

2.8 Age-period-cohort models

A particular case of modelling vital rates is the age-period-cohort model, in which the explanatory variates are all essentially demographic in nature. A table of age-specific rates for a number of time periods is constructed or abstracted from official statistics, and subjected to analysis, taking age group, time period and period of birth as the independent variables. It should be borne in mind that none of the above three factors are direct influences on mortality. Rather an age-period-cohort analysis is motivated by the idea that each is a proxy for a particular type of risk factor. Age might be considered as representing physiological variables associated with the ageing process, while period represents contemporary influences on mortality, such as environmental hazards. Cohort is usually held to stand for historical or delayed influences which are more or less common to all those with the same period of birth.

An example of an age-period-cohort analysis is provided by Duffy & Latcham (1986) who analysed liver cirrhosis mortality in Scotland and England & Wales over the period 1941-1981 in five year intervals by sex and age-group in five-year bands from 30-34 to 70-74.

Table 10 contains a subset of the data for Scottish males for the period 1961-1981 taken from the larger table in the original paper.

Table 10: Liver cirrhosis mortality in Scotland 1961-1981

Age	1961		1966		1971		1976		1981	
Group	r	n	r	n	r	n	r	n	r	n
30-34	3	163	0	154	1	147	1	157	4	185
35-39	2	170	1	156	4	147	4	145	10	155
40-44	2	154	6	163	3	151	8	143	13	144
45-49	7	163	14	148	9	156	20	146	30	142
50-54	19	165	16	154	12	140	25	149	36	142
55-59	21	148	23	152	13	142	24	131	45	141
60-64	23	114	24	130	18	134	30	127	45	118
65-69	20	84	25	93	25	107	32	111	35	108
70-74	13	60	13	62	16	69	21	79	21	86

The columns headed r contain the numbers of deaths in the appropriate age group and period, and the corresponding total population at risk (in thousands) is denoted by n. The table therefore contains information concerning five time periods and nine age-bands. The basis of cohort identification is that the population at risk of death in age-band 30-34 in 1961 comprises the same individuals (apart from losses by death or migration) as that in age-band 35-39 in 1966 and so on. Thus the data represents the experience of several birth cohorts, easily deduced to be 13 in all.

However the level of the cohort factor for a particular r and n is a linear combination of the corresponding levels of age and period, satisfying the relation

$$\mathrm{COH} = a + \mathrm{PER} - \mathrm{AGE} \qquad (4.1.1)$$

where COH, PER and AGE represent the levels of these factors, and a is the total number of levels of AGE. Thus in this formulation cohort 1 for the above data were the oldest age group in 1941, and appear only once in the dataset. Cohort 13 corresponds to the youngest age group in 1981 and also appears only once. This linear dependence leads to the so-called identification problem. The linear effect of cohort is confounded with the linear effects of period and age, or more precisely all three effects are confounded. This leads to unidentifiability of one of the linear effects in the model, and implies that for estimation of the full model, fitting each of the factors together there are

$$a+p+c-4 \qquad (4.1.2)$$

degrees of freedom, one less than might be expected, where a, p and c are the number of levels of the respective factors.

GLIM example 9

To illustrate this method the following GLIM analysis was performed on the data.

```
$UNITS 45
$DATA R N
$DINPUT 10
FILE NAME? COHORTDAT
```

The data are read in from a file containing only numbers of deaths and numbers at risk, with age and period calculated within GLIM by means of the %GL function.

```
$CAL PER=%GL(5,1)
$CAL AGE=%GL(9,5)
$
```

These instructions assign values from 1 to 5 in sets of 1 to the vector PER, and values from 1 to 9 in sets of 5 to AGE. The vectors themselves are of length 45, set by the $UNITS directive. The values correspond to data being entered by periods within age groups.

```
$CAL N=N*1000
$YVAR R
$ERROR B N
$FAC AGE 9 PER 5
$FIT AGE+PER$

scaled deviance = 36.616 at cycle  3
            d.f. = 32
```

We may now create and declare the cohort variable as a factor at 13 levels, as follows

```
$CAL COH=9+PER-AGE
$FAC COH 13$
$FIT +COH$
scaled deviance = 17.021 (change =  -19.59) at cycle
4
            d.f. = 21     (change =  -11    )
```

Notice that there are only 11 degrees of freedom for the cohort factor, despite it having 13 levels. this illustrates the confounding of the linear effect of cohort with the other two factors. The change in deviance shows that terms in cohort are not required in the model, that is that there are no significant non-linearities associated with the levels of cohort. For illustrative purposes it is nevertheless useful to show the effect estimates.

```
$DIS E$

     -11.68         0.3910       1
      1.060         0.4681       AGE(2)
      1.649         0.4503       AGE(3)
      2.681         0.4153       AGE(4)
      3.069         0.4023       AGE(5)
      3.255         0.3988       AGE(6)
      3.453         0.4024       AGE(7)
      3.569         0.4149       AGE(8)
      3.246         0.4226       AGE(9)
     0.1783         0.1467       PER(2)
    0.04866         0.1886       PER(3)
     0.5912         0.2257       PER(4)
     0.9416         0.2638       PER(5)
    -0.2216         0.3293       COH(2)
    -0.2146         0.3135       COH(3)
    -0.4183         0.3170       COH(4)
    -0.6427         0.3329       COH(5)
    -0.8230         0.3707       COH(6)
    -0.6591         0.4107       COH(7)
    -0.6134         0.4540       COH(8)
    -0.5698         0.4995       COH(9)
    -0.3807         0.5506       COH(10)
    -0.2820         0.6191       COH(11)
   -0.09225         0.6612       COH(12)
      0.000        aliased       COH(13)
```

Again, the confounding is illustrated by the absence of an estimate for cohort 13. The next question of interest is to assess whether both age and period are needed in the model. We may address this by deleting them in turn, fitting models in age only and period only, and also assess whether either of the factors may be replaced by a variate. In this instance it turns out that the answer in both cases is that they cannot, as may be seen by comparing the residual deviances from the two models below with that for the first model fitted. Accordingly it follows that both factors are needed.

```
$VAR AGE
 -- change to data affects model
$FIT PER+AGE$
 scaled deviance = 104.75 at cycle  4
             d.f. =  39
$VAR PER
 -- change to data affects model
```

```
$FAC AGE  9
$FIT AGE+PER$
 scaled deviance = 57.844 at cycle  3
            d.f. = 35
```

It remains only to display the estimates from the model including both age and period as factors.

```
$DIS E$
estimate        s.e.      parameter
  -4.854      0.3475      1
  0.9425      0.4006      AGE(2)
   1.416      0.3798      AGE(3)
   2.416      0.3545      AGE(4)
   2.766      0.3500      AGE(5)
   3.013      0.3484      AGE(6)
   3.325      0.3479      AGE(7)
   3.557      0.3491      AGE(8)
   3.329      0.3571      AGE(9)
 0.08793      0.1424      PER(2)
 -0.1537      0.1483      PER(3)
  0.4368      0.1351      PER(4)
  0.9322      0.1285      PER(5)
```

It can be seen that the risk of cirrhosis increases up to age group 8, and declines slightly for the oldest group (70-74). Periods 4 and 5 are associated with larger risks than period 1, 2 and 3.

2.9 A note on the identification problem

Several approaches have been suggested to overcome the identification problem, and a useful summary of the subject is given by Fienberg & Mason (1985). A method suggested by them is to use an identification specification, that is to assume at the outset that one or more effect estimates for particular levels of one of the factors are equal. This certainly permits estimation of the other parameters, but the magnitudes of the resulting estimates depend on the identification specification used. An identification specification is easily implemented in GLIM by defining the levels of the factor which are assumed to have equal effects (the equality set) to be the same, usually the lowest of the levels in the set.

More recently, Clayton & Schifflers (1987) in a comprehensive review of the age-period-cohort problem have demonstrated that no matter how the model is cast an arbitrary linear trend may be added to one set of estimates, making concomitant adjustments to the others, without loss of a degree of freedom. This "drift" component cannot be apportioned uniquely to any of the three categories. They conclude that use of the methodology be restricted to descriptive studies.

It is possible to agree with Clayton's point without necessarily discounting the value of models of this type. It is sometimes the case that only two of the explanatory factors are required in the model, and it would seem perverse in those circumstances to refuse to make any inference simply because an arbitrary linear trend in the third could be accommodated in the model without changing either the adequacy of explanation (in terms of goodness of fit) or the level of complexity of the model (the degrees of freedom). If there is no necessity to assume a non-zero trend in the third factor, then it is reasonable to assume it is zero.

Another approach to the identification problem is to substitute an external variate for one of the classifications. In an age-period-cohort analysis of suicide in England & Wales, Duffy & Surtees (1989) substituted the level of carbon monoxide in domestic gas for the period component and obtained a good fit to their data for males. It must be conceded that an arbitrary linear trend in period might be added to the model, but this would once again be involving a variate which does not need to appear.

III. SURVIVAL ANALYSIS IN GLIM

3 Survival analysis in GLIM.

Although the main focus of this course is the use of logistic-linear models in epidemiological research, the GLIM package may also be used to perform the analysis of survival data from clinical trials and other epidemiological studies. Some simple methods will be illustrated, and attention will be directed mainly towards implementation of the methodology rather than detailed description of theory. Other statistical packages such as BMDP, SPSS-X etc also have routines to perform these types of analyses.

The implementation of methods of survival analysis in GLIM relies for the most part on specially written routines or sets of GLIM instructions which are called MACROs. These may be stored on the computer system in question and read into a GLIM program as required using the $INPUT directive. In effect the GLIM system is being as a programming language or a calculator rather than a model-fitting system for some of these applications.

3.1 Basic concepts

Survival analysis is concerned with statistical description of the process governing times to death or some other key event of individuals in an epidemiological study. In most situations we are concerned to compare two or more groups, to assess for example whether a particular treatment or exposure to a risk factor changes the expected time to death. Various characteristics of the study members may be recorded, and their influence on survival adjusted for in group comparisons.

It is not usually practicable to follow up all subjects selected for study until the occurrence of the event of interest. Individuals may be lost to follow-up through withdrawal of cooperation, because of geographical moves, or for other reasons. Standard practice is to study the groups for a fixed period of time, and not all individuals will have experienced the event of interest at the end of this period. Thus, considering the outcome for a study member to be time from an appropriate initial point (such as date on which an illness was diagnosed) to time of the event, for some cases the information available is only that the event had not occurred at time t_i, and we say that the observation is censored (or right-censored, as it sometimes happens that we may know the time of the event, but not the appropriate start time for an individual, in which case the observation is left-censored). This type of censoring is often called progressive censoring, to distinguish it from the situation in which all individuals have the same starting time in a study of fixed length (single censoring), as in carcinogenicity experiments on animals. Methods of dealing with censoring described in this chapter assume that censoring occurs at random, and is not related directly to group membership or subsequent survival.

Of the possible approaches to this problem it is worth making a distinction from the start between parametric or distributional methods and distribution-free methods. The first type involve the assumption of a particular probability distribution governing times to death, and estimation of its parameters, whereas distribution-free approaches do not require the assumption of a particular distributional form.

3.2 Introductory theory

Understanding ofthe methods to be illustrated will be facilitated by knowledge of the statistical basis of the subject. In this section some elementary ideas of survival analysis will be described mathematically, and for simplicity of terminology we shall consider the event of interest to be death.

We may describe the probability distribution of time to death from an appropriate initial point in a number of ways. The cumulative distribution function of time to death, T, that is the probability of dying before t, is defined as

$$F(t) = P[\,T \leq t\,] \qquad : t \geq 0 \quad (3.2.1)$$

and from this we may define the death density function assuming F(t) to be differentiable as

$$f(t)=\frac{dF(t)}{dt} \qquad (3.2.2)$$

However it is more usual to consider the survival function, the probability that an individual survives until at least time t, which is just the complement of F(t)

$$S(t) = 1\text{-}F(t) \qquad : \qquad t \geq 0 \quad (3.2.3)$$

A function related to these and very widely used is the hazard function, also called the instantaneous death rate, failure rate or force of mortality. It may be denoted by λ(t), and the probability of death in the infinitesimal interval (t,t+δt), given survival to time t is given by

$$\lambda(t)\ \delta t \quad (3.2.4)$$

It is easy to show that λ(t) is related to the previously defined quantities as

$$\lambda(t) = \frac{f(t)}{S(t)} \qquad (3.2.5)$$

Several explicit functional forms are commonly applied in the analysis of survival data. For example a constant hazard rate leads to the exponential distribution of times to death. This may be generalised to a gamma distribution, which permits the hazard rate to increase or decrease with time according to whether the shape parameter is > or < 1. The corresponding hazard function for integer n is

$$\lambda(t)=\frac{\lambda^n t^{n-1} / (n-1)!}{\sum_{v=0}^{n-1} (\lambda t)^v / v!} \quad : \lambda,t \geq 0 \qquad (3.2.6)$$

The hazard rate

$$\lambda(t) = \lambda\gamma\,(t\text{-}\delta)^{\gamma\text{-}1} \qquad : \lambda,\gamma,\delta \geq 0 \qquad (3.2.7)$$

leads to the Weibull distribution of times to death. This hazard rate is widely used in the analysis of survival data, often with δ assumed known or zero. As in the case of the gamma model, the Weibull hazard increases or decreases with time according as γ is greater or less than 1, and provides a good fit to survival data of many types. Both the exponential and Weibull distributions may be fitted to censored survival data using a macro, originally described by Aitkin & Clayton(1980) supplied in the GLIM macro library.

3.3 Lifetables and the Kaplan-Meier survival estimate.

A set of survival data may be simply described by construction of a lifetable, and graphically represented by a survival curve. Data may be grouped into intervals by time of deaths, or considered as a set of individual observations. An example of the latter is provided by the survival times of 38 patients diagnosed as suffering from adenoid cystic carcinoma reported by Marsh and Allen, 1979. Appendix C summarises selected variates from those published.

Times in column 1 are years from onset of symptoms to diagnosis. Variables in columns 2 and 3 indicate methods of treatment used, 1 corresponding to use of the method, 0 to those cases for which the method was not used. Zero in column 4 indicates that the patient was still alive at the time of ending data collection, that is the observation is censored, and a 1 that the time in column 5 is indeed a time to death (in months).

For samples as small as this the Kaplan-Meier (1958) approach is particularly suitable. Times of death, t_i, are ordered from smallest to largest, and for any particular $t_{(i)}$ we may write the number of patients still at risk as n_i, which includes the patient dying at t_i. If one of the censored times is also equal to t_i the Kaplan-Meier method includes the censored patient in the n_i. The probability of surviving at least time t (between values $t_{(i)}$ and $t_{(i+1)}$ say) is estimated by

$$\hat{S}(t) = \prod_{j=1}^{i} \frac{n_j - 1}{n_j} \quad (3.3.1)$$

If there is no censoring then this estimate reduces to the simple proportion of survivors at time t.

An approximate estimate of the variance of $\hat{S}$ is given by

$$\hat{\text{var}}[\hat{S}(t)] \cong [\hat{S}(t)]^2 \sum_{j=1}^{i} \frac{1}{n_j(n_j-1)} \qquad (3.3.2)$$

Estimation of the survival function by this method is a computational rather than a modelling task. However the GLIM package can be used to perform the calculations involved using specially written macros. Swan (1986) published a macro which estimates the survival function, produces a line-printer plot of the function, and will also perform a logrank test of significance of the difference between two survival distributions.

The survival macro as implemented for the purpose of this text is named SURV, but is otherwise identical to that originally published. It takes five arguments, the times on study, a censoring indicator, a weight variable, which determines whether a case is used or not, an indicator which determines how the intervals to be used are constructed and another which permits censored observations to be dealt with in different ways.

GLIM example 10

To apply this to the data in appendix C requires the following control language

```
$UNITS 38
```

```
$DATA SYMT SURG RT IND TIME
$READ followed by the data, or alternatively
$DINPUT 10 followed by the appropriate filename
$CAL W=1 which indicates that all the data are to be used
$CAL %K=0
$CAL %L=1
$INPUT 11 followed by the filename containing the macro
$USE SURV TIME IND W %K %L
```

The value of the fourth parameter %K in this case is 0, indicating that intervals in which no death occurred may be omitted, while the fifth parameter, set to 1 indicates that censored observations should be considered at risk for the whole of the intervals in which loss occurs.

The survival estimates and standard errors are given in table 10, and the corresponding lineprinter graph plotted as figure 2.

Separate analyses may be performed for subgroups of the data. One might for example wish to estimate different survival curves depending on treatment by radiotherapy. This is easily implemented by changing the weights W by

```
$CAL W=(RT==1)
```

to select those who did receive radiotherapy and exclude those who did not, and rerunning the macro.

Table 11:Lifetable with estimated survival function and se

	Top of	No. at	deaths	Lost	Hazd		Est p	
	Intvl	Risk			d/(n-fw)	1-q	Surv	
	t	n	d	w	q	p	S	se
	TI_	NI_	DI_	WI_	QI_	PI_	SI_	SES_
1	0.0	38.0	0.0	0.0	0.000	1.00	1.000	0.000
2	1.0	38.0	1.0	0.0	0.026	0.97	0.974	0.026
3	2.0	37.0	0.0	1.0	0.000	1.00	0.974	0.026
4	16.0	36.0	1.0	0.0	0.028	0.97	0.947	0.037
5	19.0	35.0	1.0	0.0	0.029	0.97	0.920	0.045
6	23.0	34.0	1.0	0.0	0.029	0.97	0.893	0.051
7	28.0	33.0	1.0	0.0	0.030	0.97	0.865	0.056
8	33.0	32.0	1.0	0.0	0.031	0.97	0.838	0.060
9	35.0	31.0	0.0	1.0	0.000	1.00	0.838	0.060
10	37.0	30.0	1.0	0.0	0.033	0.97	0.811	0.065
11	39.0	29.0	0.0	2.0	0.000	1.00	0.811	0.065
12	40.0	27.0	1.0	0.0	0.037	0.96	0.780	0.069
13	49.0	26.0	1.0	0.0	0.038	0.96	0.750	0.072
14	55.0	25.0	1.0	0.0	0.040	0.96	0.720	0.075
15	57.0	24.0	1.0	0.0	0.042	0.96	0.690	0.078
16	58.0	23.0	0.0	1.0	0.000	1.00	0.690	0.078
17	60.0	22.0	1.0	0.0	0.045	0.95	0.659	0.081
18	67.0	21.0	1.0	0.0	0.048	0.95	0.628	0.083
19	78.0	20.0	0.0	1.0	0.000	1.00	0.628	0.083
20	79.0	19.0	1.0	0.0	0.053	0.95	0.595	0.085
21	81.0	18.0	2.0	0.0	0.111	0.89	0.529	0.087
22	84.0	16.0	1.0	0.0	0.063	0.94	0.496	0.088
23	85.0	15.0	1.0	0.0	0.067	0.93	0.462	0.088
24	87.0	14.0	1.0	0.0	0.071	0.93	0.429	0.088
25	91.0	13.0	1.0	0.0	0.077	0.92	0.396	0.087
26	106.0	12.0	1.0	0.0	0.083	0.92	0.363	0.086
27	108.0	11.0	0.0	1.0	0.000	1.00	0.363	0.086
28	117.0	10.0	1.0	0.0	0.100	0.90	0.327	0.084
29	123.0	9.0	0.0	1.0	0.000	1.00	0.327	0.084
30	126.0	8.0	1.0	0.0	0.125	0.88	0.286	0.083
31	127.0	7.0	1.0	0.0	0.143	0.86	0.245	0.081
32	132.0	6.0	1.0	0.0	0.167	0.83	0.204	0.077
33	157.0	5.0	1.0	0.0	0.200	0.80	0.164	0.072
34	168.0	4.0	0.0	1.0	0.000	1.00	0.164	0.072
35	174.0	3.0	0.0	1.0	0.000	1.00	0.164	0.072
36	211.0	2.0	1.0	0.0	0.500	0.50	0.082	0.068
37	236.0	1.0	1.0	0.0	1.000	0.00	0.082	0.068

Figure 2: Lifetable survival curve

```
   1.0000 +
   0.9474 2  +
   0.8947 |   ++
   0.8421 |     +2
   0.7895 |      +2
   0.7368 |         ++
   0.6842 |          +2
   0.6316 |            +  +
   0.5789 |               +
   0.5263 |               +
   0.4737 |                2
   0.4211 |                ++
   0.3684 |                    ++
   0.3158 |                      + +
   0.2632 |                        2
   0.2105 |                         +
   0.1579 |                              +  ++
   0.1053 |                                         +    +
   0.0526 |
   0.0000 |
----------:---------:---------:---------:---------:---------:----:
    '    0.0      50.0     100.0     150.0     200.0     250.0

 Estimated median survival time=    83.59   with se=    10.70
```

Figures 3 and 4 show the estimated survival function graphs for the two groups. The median survival times are not very different but the shapes of the survival curves are quite dissimilar, although it should be noted that these are drawn on different x scales. It can be seen that survival estimates do not appear to change after about 100 days for the radiotherapy treated patients. However the numbers of patients involved are very small, quite apart from the other aspects of the patients and their experience which must be taken into account before attempting to draw any conclusions from an analysis.

Figure 3: Lifetable survival curve- radiotherapy

```
   1.0000 ++
   0.9474 |      +
   0.8947 |       ++
   0.8421 |
   0.7895 |        ++
   0.7368 |           +
   0.6842 |             ++
   0.6316 |
   0.5789 |                +
   0.5263 |
   0.4737 |                   +
   0.4211 |                    +
   0.3684 |
   0.3158 |                     +    +   +          + +
   0.2632 |
   0.2105 |
   0.1579 |
   0.1053 |
   0.0526 |
   0.0000 |
----------:---------:---------:---------:---------:---------:------
       0.0      40.0      80.0     120.0     160.0     200.0

Estimated median survival time=   80.89  with se=   4.576
```

Figure 4: Lifetable survival curve- no radiotherapy

```
   1.0000 +
   0.9474 +
   0.8947 |  +
   0.8421 |   +
   0.7895 |    +  +
   0.7368 |       +
   0.6842 |          +
   0.6316 |           +   +
   0.5789 |               +
   0.5263 |               +
   0.4737 |                +
   0.4211 |                 +
   0.3684 |                    +
   0.3158 |                      +
   0.2632 |                        +
   0.2105 |                        +
   0.1579 |                         +
   0.1053 |                              +
   0.0526 |                                         +    +
   0.0000 |
----------:---------:---------:---------:---------:---------:------
        0.0       50.0     100.0     150.0     200.0     250.0

Estimated median survival time=   84.88  with se=   2.885
```

3.4 The logrank test

A number of distribution-free tests have been adapted for the analysis of survival data. For the case in which there is no censoring the appropriate nonparametric test may be applied without modification to the times of death. For example two groups may be compared by means of the Wilcoxon rank sum test, or equivalently the Mann-Whitney U-test, while an appropriate test for several groups would be the Kruskal-Wallis analysis of variance by ranks. Censoring however necessitates adaptation of the tests.

The Wilcoxon test may be adapted following Gehan(1965) by considering all possible pairs of observations, containing one member from each sample. In calculating the test statistic using the Mann-Whitney U calculation pairs of censored observations do not contribute to the total, neither do cases in which the smaller member of the pair is censored, but all other pairs are counted in the usual fashion. Assuming random censoring the permutation distribution of the statistic may be constructed for statistical inference, or a normal approximation may be used.

Perhaps the most widely used distribution-free test for this problem is the logrank test (Peto & Peto,1972). The test is applicable to the comparison of any number of groups, but is most simply described for the two group case. Only intervals in which one or more deaths occur are involved in the calculation, which essentially involves dividing the observed number of deaths in a particular interval into expected numbers of deaths in each in accordance with the number of individuals at risk in the groups. Thus if say on a particular day i there are d_i deaths, and n_{1i} and n_{2i} persons at risk in each group the expected number of deaths in the first group is given by

$$\frac{d_i\, n_{1i}}{n_{1i}+n_{2i}} \qquad (3.4.1)$$

while those for the second group are given by (3.4.1) replacing n_{1i} in the numerator by n_{2i}.

The total expected numbers of deaths for each group, E_1 and E_2 say are formed by adding over all the days in question, and these are compared with the total observed deaths O_1 and O_2 by calculating

$$\frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} \qquad (3.4.2)$$

which may be compared with the χ^2 distribution on 1 degree of freedom to assess the significance of the difference. When more than two groups are involved formula 3.4.2 is simply extended to the appropriate number of groups, and the degrees of freedom for the test are increased appropriately.

The death rate ratio between the two groups may be estimated as

$$\frac{O_1/E_1}{O_2/E_2} \qquad (3.4.3)$$

In cases where it is not reasonable to assume constant death rates within each group over time 3.4.3 is effectively an average death rate ratio.

The paper by Swan (1985) provides a macro for performing the logrank test for two groups, taking as input the vectors constructed by the macro SURV applied to each group. The total numbers of deaths and subjects at risk are contained in the vectors di_ and ni_ after using the macro. The procedure therefore is to define a weight variable to be 1 for the first group, and 0 for the second, use the macro, rename the above vectors, redefine the weight variable to select the second group, use the macro again and then rename the vectors and use the macro LRT.

GLIM example 11

Assuming that the adenoid cystic carcinoma data has been entered into GLIM as for the previous example we might compare death rates between those who received radiotherapy and the remainder by the logrank test as follows

```
$USE SURV TIME IND RT %K %L
$CAL NR=NI_
$CAL DR=DI_
$CAL W=%IF(RT==0,1,0)
$USE SURV TIME IND RT %K %L
$CAL ND=NI_
$CAL DD=DI_
$USE LRT NR DR ND DD
```

The output from the above is particularly simple

```
 The logrank chisquared on 1 df is   0.1490
```

Extension to comparisons of several groups is straightforward by making minor modifications to the macro.

3.5 Cox regression and the proportional hazards model

Equation 3.2.4 defined the hazard rate in terms of the probability of death in the interval $(t,t+\delta t)$ given survival to time t. Cox (1972) suggested use of partial likelihood methods based on hazard rates of the form

$$\lambda(t,\mathbf{x}) = \lambda_o(t)\exp(\beta'\mathbf{x})(3.5.1)$$

where $\lambda_0(t)$ is an arbitrary unknown function of t, **x** is a vector of explanatory variables and β is a vector of unknown parameters.

Considering a death of an individual with explanatory variables $\mathbf{x}_i$ occurring at t_i, the contribution to the overall likelihood may by a conditional argument be written as

$$\frac{\exp(\beta'\mathbf{x}_i)}{\sum_j \exp(\beta'\mathbf{x}_j)} \qquad (3.5.2)$$

where the suffix j runs over all those individuals alive and not censored at time t_i, that is those at risk of death at t_i. Notice that $\lambda_0(t)$ cancels, and therefore does not require to be estimated in this formulation. In the event of censoring occurring at t_i the individual concerned is counted as being at risk at t_i, as was the case for the logrank test and the Kaplan-Meier estimate.

Thus, the overall partial likelihood associated with the set of observations is the product of terms 3.5.2 over all deaths. There has been considerable theoretical investigation of this model, and among other results it has been shown that maximum likelihood estimation based on the product of terms like 3.5.2 leads to the same estimate of β as is obtained by simultaneous maximum likelihood estimation of $\lambda_0(t)$ and β.

This method is implemented on several statistical analysis systems, but attention is restricted here to the GLIM package. An initial approach to the use of GLIM for the Cox model was described by Whitehead (1980), but is sufficiently complicated to lead to suggestions that it would be simpler and more economical of computer time to write a Fortran program to perform the analysis (Morris, 1985). A more recent algorithm, easily implemented on GLIM was provided by Clayton & Cuzick (1985) which can be shown to lead to correct estimates of the unknown parameters and appropriate deviance statistics, although standard errors of the parameters are slightly underestimated.

GLIM example 12

Appendix D contains data from a random sample of 108 admissions to the Regional Poisoning Treatment Centre of the Royal Infirmary of Edinburgh following an act of self-poisoning or self-injury (parasuicide) in the years 1982 to 1985 inclusive. for each selected admission the time to the next admission for parasuicide was recorded, along with some demographic information. Most of the individuals had no subsequent admission, so their times are right-censored. The table contains information on sex, age, censoring and time to next admission in days. Males are coded as 1 on the sex variable, females as 2, ages are in years and a value 0 of the indicator marks a censored observation.

The GLIM macros to perform the analysis are identical to those given by Clayton & Cuzick (1985) and have been given the same names as in that paper. Assuming that the data in Appendix 4 has been read into GLIM with names SEX, AGE, IND and TIME the following GLIM control language is used to perform the analysis.

```
  $CAL TIME=TIME-IND*0.001
  $ARGUMENT COX1 TIME IND
  $FACTOR SEX 2
  $USE COX1
  $
  Partial likelihood deviance for null model=
254.04704
  $FIT SEX$
  scaled deviance = 119.57 at cycle  5
             d.f. = 106

  $USE COX2
  $
 -- change to data affects model
  scaled deviance = 120.23 at cycle  5
             d.f. = 106

  -- change to data affects model
  scaled deviance = 120.24 at cycle  5
             d.f. = 106
```

```
partial likelihood deviance=    252.08716
```

The above instructions first of all ensure that all observations censored at the same time as the occurrence of a death are entered into the appropriate risk set, by slightly reducing the times of deaths only. The arguments for the macro are declared to be times to repetition and the censoring indicators by the $ARGUMENT directive and the macro COX1 is used to perform an initialisation and a null fit. Then a term for SEX is introduced into the model, and the macro COX2 used to estimate the corresponding parameter and the associated deviance. The GLIM deviances are of no relevance to the analysis, and the macros provide correct partial likelihood deviances which should be used for inference. The difference in deviance between the null fit and the model including sex is about 1.96, on 1 degree of freedom, and therefore not statistically significant.

Age and age and sex together may now be fitted in the same way. Again only the partial likelihood deviances should be used for inference.

```
$FIT AGE$
 scaled deviance = 122.19 at cycle  5
            d.f. = 106

 $USE COX2
 $
 -- change to data affects model
 scaled deviance = 121.51 at cycle  5
            d.f. = 106

 -- change to data affects model
 scaled deviance = 121.51 at cycle  5
            d.f. = 106

 partial likelihood deviance=    254.03711
 $FIT SEX+AGE$
 scaled deviance = 119.56 at cycle  5
            d.f. = 105

 $USE COX2$
 -- change to data affects model
 scaled deviance = 120.23 at cycle  5
            d.f. = 105

 -- change to data affects model
 scaled deviance = 120.23 at cycle  5
            d.f. = 105

 partial likelihood deviance=    252.08583
 $DISPLAY E$
          estimate        s.e.     parameter
     1     0.2454       0.4972     1
```

```
2     -0.5221       0.3717      SEX(2)
3   0.0004776      0.01258      AGE
scale parameter taken as  1.000
```

The parameter estimates suggest that men are at greater (albeit not significantly greater) risk of repetition, but that age appears to have no appreciable influence on risk.

The standard errors of the parameters are likely to be close to those that would be obtained by an exact analysis, but Clayton & Cuzick suggest obtaining a confidence interval for a parameter by trial and error substitution of fixed values of the parameter into the model and examining the resulting partial likelihood deviances. This substitution may be achieved by use of the GLIM $OFFSET directive entering the fixed value of the parameter, and then maximisation using the MACRO over the remaining parameter values. The deviance changes are assessed by comparison with the appropriate %age points of the χ^2 distribution.

BIBLIOGRAPHY

AITKIN, M. & CLAYTON, D. (1980) *The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM.* Journal of the Royal Statistical Society, Series C 29: 156-163.

BRESLOW, N.E. & DAY, N.E. (1980) *Statistical Methods in Cancer Research Volume 1.* International Agency for Research on Cancer, Lyon.

BROWN, G.W. & HARRIS, T. (1978) *Social origins of depression: a reply.* Psychological Medicine 8: 577-588

CLAYTON, D. & CUZICK, J. (1985) *The EM algorithm for Cox's regression model using GLIM.* Journal of the Royal Statistical Society, Series C 34: 148-156

CLAYTON, D. & SCHIFFLERS, E. (1987) *Models for temporal variation in cancer rates. II. Age-period-cohort models.* Statistics in Medicine 6: 469-481.

CLAYTON R.M., CUTHBERT, J., PHILLIPS, C.I., BARTHOLOMEW, R.S., STOKOE, N.L., FFYTCHE, T., REID, J.MCK., DUFFY, J., SETH, J. & ALEXANDER, M. *Analysis of individual cataract patients and their lenses- a progress report.* Experimental Eye Research 31: 553-566.

CORNFIELD, J. (1951) *A Method of estimating comparative rates from clinical data.* Journal of the National Cancer Institute 11: 1269-1275.

COX, D.R. (1972) *Regression models and life tables.* Journal of the Royal Statistical Society, Series B 34: 187-203

DANIEL, C.& WOOD, F.S. (1971) *Fitting equations to data.* Wiley: New York

EVERITT, B.S. & SMITH A.M.R. (1979) *Interactions in contingency tables: a brief discussion if alternative definitions.* Psychological Medicine 9: 581-583.

FAREWELL; V.T. (1979) *Some results on the estimation of logistic models based on retrospective data.* Biometrika 66:27-32.

GART J.J. & THOMAS D.G. (1972) *Numerical results on approximate confidence limits for the odds ratio.* Journal of the Royal Statistical Society, Series B 34: 441-447.

GART, J.J. (1979) *Statistical analyses of the relative risk.* Environmental Health Perspectives 32: 157-167.

GEHAN, E.A. (1965) *A generalized Wilcoxon test for comparing arbitrarily singly-censored samples.* Biometrika 52: 202-223.

HEALY, M.J.R. (1988) *GLIM - An Introduction.* Oxford University Press, Oxford.

KAPLAN, E.L. & MEIER, P. (1958) *Nonparametric estimation from incomplete observations.* Journal of the American Statistical Association 53: 457-481.

MARMOT, M.G., SHIPLEY, M.J., ROSE, G. & THOMAS, B.J. (1981) *Alcohol and mortality- a U-shaped curve.* Lancet, 14 March: 580-583.

MARSH, W.L. & ALLEN, M.S. (1979) *Adenoid cystic carcinoma- biologic behaviour in 38 patients.* Cancer 43: 1463-1473.

MCCULLAGH, P. & NELDER, J.A. (1983) *Generalized Linear Models*. Chapman & Hall, London.

MIETTINEN, O.S. (1976) *Estimability and estimation in case-referent studies*. American Journal of Epidemiology 103: 226-235.

MORRIS, R.W. (1985) *On the application of Cox's proportional hazards model in GLIM* . GLIM Newsletter 9: 35-36

PAYNE, C.D. (ed) (1985) *The GLIM System Release 3.77- Manual*. NAg Ltd: Oxford.

PEQUIGNOT,G., TUYNS, A.J. & BERTA,J.L. (1978) *Ascitic cirrhosis in relation to alcohol consumption*. International Journal of Epidemiology 7: 113-120.

PETO, R. & PETO, J. (1972) *Asymptotically efficient rank invariant test procedures*. Journal of the Royal Statistical Society, Series A 135: 185-207.

ROTHMAN, K.J. (1986) *Modern Epidemiology*. Little, Brown & Co: Boston.

SILVA, L.C. (1987) *Ikerketa Epidemologikorako Estatistika-Azterpideak*. Eustat: Vitoria-Gasteiz

SURTEES, P.G. & DUFFY, J.C. (1989) *Suicide in England & Wales 1946-1985: an age-period-cohort analysis*. Acta Psychiatrica Scandinavica (in press)

SWAN, A.V. (1988) *Lifetables and survival curves in GLIM*. GLIM Newsletter 16: 21-27.

SWAN, A.V. (1985) *GLIM 3.77- Introductory Guide*. NAg Ltd: Oxford.

WALTER, S.W. & HOLFORD, T.R. (1978) *Additive, multiplicative and other models for disease risks*. American Journal of Epidemiology 108: 341-346.

WHITEHEAD, J. (1980) *Fitting Cox's regression model to survival data using GLIM*. Journal of the Royal Statistical Society, Series C 29: 268-275.

WILLIAMS, D.A. (1987) *Generalized linear model diagnostics using the deviance and single case deletions*. Journal of the Royal Statistical Society, Series C 36:181-191.

Appendix A: Drinking days, weekly consumption and experience of alcohol dependence symptoms of 223 company directors

DS	WKU	AP	DS	WKU	AP	DS	WKU	AP
3	15	0	7	50	1	2	16	0
3	24	0	4	29	1	1	4	0
6	82	1	2	5	0	1	12	0
7	65	0	7	210	1	4	32	0
3	23	0	1	3	0	5	33	0
4	48	1	4	18	0	7	50	0
4	29	0	7	81	0	1	14	0
3	13	0	3	20	0	1	1	0
3	19	0	7	59	0	5	26	0
2	24	0	2	12	0	6	46	1
1	1	0	7	78	0	1	4	0
5	25	0	6	56	0	5	16	0
6	129	1	3	29	0	7	54	0
3	28	0	5	40	0	1	4	0
3	45	1	7	126	0	3	14	0
2	9	0	7	52	0	4	57	1
2	52	0	6	30	0	4	84	0
2	13	0	1	28	1	2	25	0
4	48	1	1	12	0	3	24	0
5	26	0	1	16	0	1	2	0
6	62	0	1	28	0	1	12	0
3	13	0	4	19	0	5	38	0
5	35	0	4	43	0	7	72	0
5	54	1	3	16	0	7	74	0
2	16	0	3	54	0	2	30	0
1	10	0	1	14	0	2	26	0
3	21	0	5	22	0	1	1	0
5	38	0	6	110	0	5	38	0
5	42	0	1	13	0	7	74	0
5	40	0	6	92	1	5	64	0
2	8	0	7	92	0	4	26	0
1	14	0	2	8	0	7	97	0
2	18	0	5	39	0	1	2	0
2	17	0	4	30	0	2	32	0
7	63	0	5	40	1	4	7	0
7	201	1	7	102	1	7	30	0
1	10	1	7	33	0	7	54	0
2	22	0	1	6	0	6	38	0
3	38	0	7	108	0	3	20	0
1	1	0	6	40	0	6	14	0
6	36	0	2	16	0	6	36	0
3	11	0	7	35	0	6	61	0

DS	WKU	AP
2	9	0
1	3	0
3	13	0
3	70	0
3	20	0
3	14	0
7	162	0
3	35	0
1	6	0
5	48	0
2	11	0
5	62	0
5	49	1
4	36	1
6	73	0
2	5	0
7	190	1
2	10	0
1	4	0
3	6	0
2	11	0
2	13	0
4	28	0
7	238	1
5	71	0
7	83	0
2	12	0
3	36	0
7	47	0
6	66	0
1	6	0
4	44	0
1	14	0
1	6	0
5	185	0
3	38	0
5	72	0
5	31	0
1	3	0
2	51	1
6	35	0
7	110	1
7	84	0
7	58	1
3	24	0
2	10	0
1	2	0
1	1	0
4	17	0
3	26	0
1	11	0
6	43	0
2	35	0
3	12	0
7	132	1
6	36	0
6	30	0
3	6	0
4	16	0
5	40	0
1	3	0
1	1	0
1	15	0
1	20	0
3	20	1
5	33	0
1	2	0
7	48	0
7	71	1
1	10	0
6	55	1
3	25	0
2	17	0
4	68	1
7	156	0
6	37	0
1	8	0
6	54	0
3	28	0
4	53	0
4	25	0
2	16	0
6	24	0
1	1	0
3	23	0
7	75	0
2	8	0
7	34	0
7	68	0
6	26	0
6	76	0
7	147	0
4	55	0
5	23	0
2	6	0
3	219	0
1	10	0

Appendix B: Appendectomy data- case-control differences

EVENT	DIFF	SUPPORT
1	1	0
1	1	0
0	0	1
0	1	0
1	0	0
0	1	0
1	0	0
1	1	0
0	0	-1
1	0	0
1	-1	1
-1	0	0
1	0	0
1	1	0
0	-1	1
1	0	0
0	-1	1
0	-1	0
1	0	0
1	1	0
0	0	0
0	0	0
1	0	0
-1	0	1
0	0	-1
0	0	-1
0	0	0
1	1	0
0	0	1
0	1	-1
0	1	1
1	0	0
-1	0	-1
1	1	-1
0	-1	-1
1	0	0
1	1	-1
0	-1	1
0	0	0

Appendix C: Survival of adenoid cystic carcinoma patients

Duration of symptoms	surgery	radio therapy	censoring indicator	time on study
11.0	1	0	1	126
3.5	0	0	1	19
0.2	1	0	1	117
3.0	1	1	1	33
4.0	1	1	0	35
3.0	1	0	1	236
-1.0	1	0	1	132
0.4	1	0	1	85
0.1	1	1	0	168
1.0	1	1	0	108
3.0	1	0	0	78
5.0	1	1	0	39
0.1	1	1	0	2
0.1	1	1	0	58
0.2	1	0	1	211
0.0	1	1	0	174
-1.0	0	1	1	55
0.1	1	0	1	40
1.0	1	0	0	39
0.5	1	0	1	91
3.5	1	0	1	157
0.1	1	0	1	81
1.1	1	1	0	123
0.1	1	0	1	106
0.3	1	0	1	16
0.2	1	1	1	28
0.6	0	1	1	81
0.1	1	1	1	67
3.7	1	0	1	23
0.0	0	1	1	49
1.5	1	1	1	84
-1.0	1	0	1	127
3.0	0	1	1	87
2.0	0	0	1	60
0.1	1	1	1	37
0.4	1	0	1	57
0.1	1	0	1	79
0.2	1	0	1	1

Appendix D: Parasuicide repetition data

```
SEX AGE  IND   TIME                       SEX AGE  IND
TIME                SEX   AGE  IND    TIME
   1   78    0    282                        1    19     0
357                   2     22    1     499
   1   62    0   1235                        1    19     0
136                   2     21    0    1318
   1   59    0   1316                        1    17     0
168                   2     22    0     708
   1   57    1    104                        1    14     0
233                   2     22    1     392
   1   55    0    161                        2    75     0
1224                  2     21    0     917
   1   50    0    922                        2    71     0
545                   2     20    0     778
   1   49    0    472                        2    65     0
960                   2     19    0    1216
   1   46    1    316                        2    59     0
1387                  2     19    0    1194
   1   46    0    281                        2    58     0
933                   2     21    0     237
   1   43    0    628                        2    54     0
1152                  2     18    0    1140
   1   42    1    162                        2    53     0
681                   2     20    0     151
   1   42    1     71                        2    49     1
384                   2     17    0    1092
   1   40    0    931                        2    50     0
308                   2     19    0     210
   1   38    1     17                        2    48     0
521                   2     18    0     577
   1   48    1     50                        2    46     1
413                   2     16    0    1139
   1   36    1     39                        2    45     0
590                   2     15    1      82
   1   36    0    976                        2    44     0
329                   2     16    0     813
   1   34    0   1334                        2    43     1
88                    2     16    0     542
   1   34    1   1065                        2    40     0
1039                  2     16    0     430
   1   32    1     82                        2    39     0
1038                  2     15    0     424
   1   32    1    164                        2    41     0
402                   2     15    0     338
```

```
  1  31    0  1417                 2   37    1
90                 2   14    0     84
  1  30    0  1292                 2   37    0
668
  1  31    0   843                 2   36    1
85
  1  31    0   607                 2   34    0
1113
  1  30    0   365                 2   34    1
94
  1  29    0   710                 2   33    0
1112
  1  27    0  1101                 2   34    1
431
  1  29    0    67                 2   31    0
1343
  1  24    1     1                 2   32    0
597
  1  26    1    39                 2   30    1
5
  1  26    0   127                 2   29    0
1269
  1  25    0   357                 2   29    0
780
  1  22    1    46                 2   27    0
1181
  1  21    0  1238                 2   28    0
764
  1  22    0   887                 2   26    1
654
  1  22    0   527                 2   25    0
1328
  1  22    0    63                 2   27    0
554
  1  20    0   586                 2   26    1
291
  1  19    1   540                 2   27    0
19
  1  17    0  1458                 2   25    0
499
  1  19    0   609                 2   24    0
468
  1  20    1    17                 2   24    1
57
```