Controlling for socioeconomic confounding using regression methods

J F Bithell, S J Dutton, N M Neary, T J Vincent

Abstract

***Study objective* – To describe the advantages of using Poisson regression methods as an alternative to standardisation when computing expected numbers of disease occurrences adjusted for possible confounding factors. The problem of assessing the adequacy of model fit when the expectations are small is addressed by analytical calculations and by simulation. The method is illustrated with data from the national register of childhood tumours.**

***Design* – The tumour data are recorded in a national register.**

***Setting* – England, Scotland, and Wales.**

***Subjects* – The cases considered are all children registered with leukaemia or non-Hodgkin lymphoma under the age of 15 years between 1966–87.**

***Main results* – The methods show a significant variation of leukaemia incidence in relation to the Registrar General's standard region and a negative association with socioeconomic deprivation, as measured by the Townsend index. After allowing for these variables, the incidence seems to be reasonably homogeneous throughout the population, in the sense that the residual deviance does not seem to be much larger than would be expected by chance.**

***Conclusions* – The methods described have major advantages over standardisation in controlling for confounding, both in terms of flexibility of factor selection and assessment and also in the ability to determine whether there is residual variability of incidence after allowing for these factors.**

(*J Epidemiol Comm Health* 1995;49(Suppl 2):S15–S19)

A common requirement in epidemiology is the computation of expected values of the numbers of cases of a disease adjusted by possible confounding factors, such as measures of socioeconomic deprivation. Although the effects of the latter may be of interest in their own right, it is also frequently desirable to control for them in studying the association between disease and some other factor. In particular, these issues arise in the context of geographical epidemiology, where it is desired to assess the effect of an environmental factor unconfounded by any association between this factor and other spatially varying determinants of disease.

This report is intended to explain the methods and advantages of using suitable regression models for this adjustment instead of the more classic methods of standardisation. The arguments are illustrated by application to part of a large data set on childhood leukaemia (L) and non-Hodgkin lymphoma (NHL) recently used for the analysis of the possible risk of living near a nuclear installation.[1] Although the epidemiology of childhood tumours is different in some respects from that of adult disease, the data nevertheless exemplify several of the general methodological points. Specifically, there are a large number of geographical units with small expectations; and, secondly, analyses of this type can be expected to become more common with the advent of small area statistics and large scale data bases.

Methods

The data set considered consists of 5359 cases of L and NHL registered before the age of 5 years in England, Scotland, or Wales between 1966 and 1987. They were abstracted from the national register of childhood tumours maintained by the Childhood Cancer Research Group in Oxford. They were allocated to 9831 areal units which were a close approximation to 1981 electoral wards, formed by aggregating census tracts, which provide compatibility between the 1971 and 1981 censuses.[2]

For the calculation of expected numbers in these wards we abstracted population data from the Office of Population Censuses and Surveys (OPCS) files relating to the two censuses. Population estimates for inter-censal years were obtained by linear interpolation and extrapolation, but adjusting proportionately to annual district populations. They were then aggregated to provided total numbers of child-years at risk for the 22 year period. In addition, we obtained population data from the OPCS data files which were used to construct socioeconomic indicators, calculating the following variables:

- SV1 = % of men unemployed;
- SV2 = % of households owning no car;
- SV3 = % households not owner occupied;
- SV4 = % of households with more than one person per room.

Further details of these calculations are given in Bithell *et al*[1] and in a technical report available from the authors.

Department of Statistics, 1 South Parks Road, Oxford OX1 3TG
J F Bithell
S J Dutton
N M Neary

Childhood Cancer Research Group, Oxford OX2 6HJ
T J Vincent

Correspondence to:
Dr J F Bithell.

Table 1 Analysis of deviance for social class variables in relation to leukaemia and non-Hodgkin lymphoma 1966–87 (5359 cases aged 0–4 years)

Model	Residual deviance	df	Deviance due to last term	df
Null	8657·3	9830	—	—
SR	8634·2	9821	23·1	9
SR+SV1	8612·1	9820	22·1	1
SR+SV2	8619·1	9820	15·1	1
SR+SV3	8618·4	9820	15·8	1
SR+SV4	8611·0	9820	23·2	1
SR+TOWN	8610·6	9820	23·6	1
SR+TOWN+LSV1	8610·4	9819	0·2	1
SR+TOWN+SV2	8609·1	9819	1·5	1
SR+TOWN+SV3	8610·5	9819	0·1	1
SR+TOWN+LSV4	8608·9	9819	1·7	1

Key to variable names: SR = standard region; SV1 = percentage of men unemployed; SV2 = percentage of households owning no car; SV3 = percentage households not owner occupied; SV4 = percentage of households with more than one person per room; LSV1 = log (SV1); LSV4 = log (SV4); TOWN = Townsend index (see text).

POISSON REGRESSION

We used the data described above in a Poisson regression model which supposes that the numbers of registrations for a given age group in ward i have independent Poisson distributions with means μ_i, such that:

$$\log \mu_i = \log P_i + \sigma_{j(i)} + \beta x_i + \ldots, \qquad (1)$$

where:

- P_i is the total number of child years at risk in the given age group,
- $\sigma_{j(i)}$ is a (probably) small parameter giving the effect of being in the Registrar General's standard region $j(i)$;
- x_i is the value of a sociodemographic variable;
- β is a regression parameter measuring the unit effect of x_i,

and the equation can take additional similar terms to accommodate the effects of other categorical factors or quantitative variables.

This model may also be regarded as a log-linear model, which in turn is one example of a generalised linear model or GLM. The development of the theory of GLM's has provided a powerful unifying theory over the past 20 years, giving a common framework and computational methodology for analyses previously regarded as being quite different. Although the fitting of these models is more difficult than classic calculations such as standardisation, there are now numerous reliable routines in accessible statistical packages that will achieve this easily. For our calculations we used *GLIM* (*Generalised Linear Interactive Modelling*)[3], a widely available software package which permits the fitting of different models very quickly, thus simplifying the determination of which factors to take into account. This process typically involves looking at which terms lead to a statistically significant improvement in the fit of the model, as judged by changes in the residual "deviance" statistic. It should be remembered, however, that statistical significance should not be followed slavishly as a criterion for including terms; account should also be taken of the known scientific importance of the factors considered.

The parameters in such a model (eg β, $\sigma_{j(i)}$) are estimated as part of the model fitting process and represent contributions to the relative risk, thus giving a natural interpretation to the model. Thus, the difference $\sigma_3 - \sigma_2$, for example, measures the (natural) logarithm of the risk of being in standard region (SR) number 3 relative to that of being in standard region 2. Similarly, the regression coefficient measures the increase in the log relative risk associated with an increase of one unit in the associated explanatory variable. A further advantage of the modelling approach is that the so called "fitted values", obtained from the right hand side of (1) with the estimated values substituted, are precisely the ward specific expectations we require.

The general theory of Poisson regression assures us that, provided the expectations are not too small (say at least 5), the residual deviance for a model that fits adequately should have a χ^2 distribution with a number of degrees of freedom (df) equal to n, the number of observations, minus p, the number of parameters fitted. The latter includes one for the overall mean, fitting which ensures that the expected values add to the same as the observed numbers. More usefully, when we add or remove model terms, the change in the deviance also has approximately a χ^2 distribution with a number of df equal to the number of parameters added or removed. This latter property is what enables us to judge the importance of additional terms and it is also assured even when the expectations are small, provided that n is compensatingly large. By contrast, however, the residual deviance itself behaves generally quite unlike a χ^2 variate when the expectations are small, and ignoring this can give misleading conclusions about the adequacy of the fit of the model. We return to this point below, after exemplifying the modelling on the L and NHL data.

Results

Table 1 shows the result of our modelling of the numbers of children registered with L and NHL under 5 years of age, both sexes being combined. Each row represents a model with the terms specified, "null" signifying that only the overall mean is fitted. It will be seen that the residual deviances are all much smaller than the numbers of df, which would be highly unlikely if they were truly χ^2 variables determining the fit of a good model. Standard region is a regional grouping with 10 levels and consequently nine df (and independent parameters), and the deviance difference resulting from fitting this term would therefore have a χ^2 distribution with 9 df if this factor made no significant difference to the fit. The value of $\chi^2_9 = 23{\cdot}1$ is formally statistically significant ($p = 0{\cdot}006$), from which we conclude that rates do vary somewhat between regions, so that this factor is worth considering for inclusion.

Initially we added the four socioeconomic variables SV1, . . ., SV4 individually in addition to standard region; this gave deviance reductions which are all highly significant, the critical value for χ^2_1 at the 0·1% level being 10·8. We also calculated the Townsend index[4] as the following function of the four variables:

Townsend index =

$$\left(\frac{LSV1-m1}{d1}+\frac{SV2-m2}{d2}+\frac{SV3-m3}{d3}+\frac{LSV4-m4}{d4}\right)\div 4,$$

where:

- LSV1 = log (1 + SV1), LSV4 = log (1 + SV4),
- SV1–SV4 are as defined above,
- m1,d1 are the mean and SD of LSV1
- m2,d2 are the mean and SD of SV2
- m3,d3 are the mean and SD of SV3
- m4,d4 are the mean and SD of LSV4.

The means and SDs were unweighted; using the numbers of child-years at risk as weights made very little difference.

It will be seen from table 1 that the Townsend index performs about as well as unemployment and overcrowding separately and rather better than house or car ownership. In view of the fact that this index has found useful application elsewhere,[5] we decided to use it as our standard measure of socioeconomic deprivation. Adding the individual components *LSV1, SV2, SV3, LSV4* as extra terms in addition to the Townsend index produces very small further reductions in deviance, confirming that the index adequately represents the information contained in these four variables.

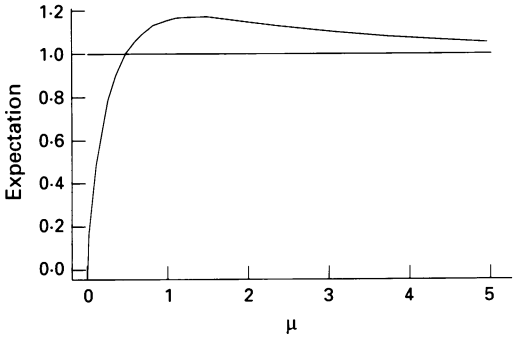
Using the exploratory modelling advantages of *GLIM*, we also investigated a number of other possible model terms. No quadratic terms for the SV variables or for the Townsend index gave a significant deviance reduction, confirming the linearity of the relationship between these indices and log relative risk. We initially followed other workers[6] and calculated the quintiles of the SV terms for use as grouping values for five level factors. The reasons for doing this are partly historical, in that standardisation involves stratification, and partly that non-linear relationships are automatically allowed for. As we found no evidence of non-linearity, we abandoned the quintile approach on the grounds that it introduces a discontinuity – that is, wards with very similar values of an SV variable near a quintile boundary lead to substantially different adjustments to the expectation. This ability to model regression like terms as opposed to factors is a further advantage of the modelling approach. As a result of the interactive modelling process, we selected as a definitive model that with SR and the Townsend index fitted.

GOODNESS OF FIT

The question of how to assess the goodness of fit of the model is not easy when – as in this case – the expectations are small. The deviance statistic returned by the fitting of a log-linear model is defined by:

$$G = 2^{*}\Sigma\{Y_i \log(Y_i/\hat{\mu}_i) - (Y_i - \hat{\mu}_i)\}$$

where Y_i is the observed count and $\hat{\mu}_i$ is the estimated expectation in ward i. We can obtain some idea of how it behaves with small expectations by supposing that, in a given term D of the sum, Y_i really does have a Poisson distribution with mean $\hat{\mu}_i$. In this case the mean E(D) of D can be computed numerically; it is shown as a function of μ in figure 1. It can be seen that, for large μ, E(D) approaches one, which would be expected from the fact that a χ^2 variate with 1 df has a mean of one. For smaller values of μ, however, E(D) can be as large as 1·160 (at $\mu = 1{\cdot}33$) or arbitrarily small (for small μ). This explains why the residual deviances observed in our data are all smaller than the number of df. The mean and variance of G calculated similarly can also be used to adjust a sum of deviance terms on the assumption that the Y_i are independent and have Poisson distributions. Table 2 shows the mean, variance, and normalised value for the null deviance for our model calculated this way; it will be seen that, before fitting the SR and Townsend score, the data show convincing evidence of an inadequate fit ($p = 0{\cdot}00042$, referring the standardised ratio to a normal distribution).



Expected value of a single contribution to the deviance statistic as a function of the expectation of the Poisson count.

A similar analysis can be performed on the expectations after fitting SR and Townsend index; it will be seen that the last column of table 2 that the fit has improved considerably ($p = 0{\cdot}025$). Although this is significant beyond the 5% level, it is indicative of a fairly low level of heterogeneity of risk, given the large size of the data set.

The problem with this analysis lies in the assumption of independence, which is not strictly true even when testing the null model, since fitting the overall mean ensures that the counts and the expectations fitted have the same sum. Similar (though more complex) constraints are implied by the fitting of each of

Table 2 Observed deviance from the null model and that with selected variables fitted, compared with their theoretical means and variances on assumptions of independence (see text).

	Null model	*SR + TOWN fitted*
Observed deviance	8657	8610·6
Degrees of freedom	9830	9820
Expectation	8330	8419·6
Variance	9571	9510·0
z-score	3·34	1·96
p value	0·00042	0·025

SR = standard region; TOWN = Townsend index.

the parameters in the definitive model and this has an effect on the mean and variance of G which is hard to evaluate, although intuition suggests that the resulting correlations between the components of G should be very weak with such a large data set. We tested the analytical conclusion by a simulation experiment in which we assumed that the $\{\mu_i\}$ are given by the definitive model in which the true parameters are actually equal to those estimated. We then simulated 500 data sets, each of size 9831, drawing observations from the Poisson distribution with means $\{\mu_i\}$. For each data set we fitted a GLM with the same terms and computed the deviance. Out of the 500 resulting deviances, 13 were greater than the value of 8610·6 actually observed. This implies that 0·026 is an unbiased estimate of the true p value for our actual data, a value in close agreement with the analytical results. The interpretation of the residual deviance by computing its theoretical mean and variance as in table 2 is therefore vindicated in this case, and it may be supposed that the method provides an attractive general way of assessing the goodness of fit of a log-linear model with small expectations. However, the method is likely to work less well with smaller data sets and it would generally be advisable to check the conclusions by simulations as we have done.

It may be mentioned that some authors recommend the use of the Pearson's χ^2 statistic instead of G in the case of small expectations. This familiar measure of discrepancy is given by:

$$K=\Sigma\frac{(Y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}$$

and is analytically much more tractable than G. Its expectation is equal to n and is therefore unaffected by the smallness of the individual area expectations. Its variance is given (on the same assumptions of independence) by $2n+\Sigma 1/\mu_i$, the latter term representing the easily computed inflation due to the small expectations. On the other hand individual terms can be very much larger than the corresponding deviance terms, particularly when $Y=1$ and μ is very small. For example, our largest component of K was 305·5 (*observed* = 1, *expected* = 0·00325), compared with 16·80 (*observed* = 7, *expected* = 0·879) for the largest component of G. Thus K is dominated by the occasional unit with a very small expectation and a single case, whereas G is more sensitive to excesses with larger counts, which would seem to give the latter a distinct interpretational advantage. The difficulties of allowing for the fitting of parameters apply equally to K and G.

Discussion

The methods we have described have resulted in the construction of a set of observed and expected numbers having considerable epidemiological value and it is our intention to use it to examine various geographical hypotheses. Childhood leukaemia is atypical of cancer in general in that it is very rare and seems to vary with explanatory variables to a substantially lesser extent than many adult cancers. It should also be noted that the association with deprivation is *negative* – that is, the risk seems to be slightly higher in higher social class groups. Nevertheless, the problem of controlling the expected values for possible confounders is very similar to that which arises in other branches of epidemiology.

The model we used assumes that the probability of disease is small and that we have large populations – the classic assumptions for the applicability of a Poisson approximation to the binomial distribution. In other circumstances it may be appropriate to use the binomial distribution instead – the model would still be a GLM and much of what we have said would apply.

We have assumed that the observed counts are independent, which seems entirely reasonable for a non-contagious disease. Some authors[7,8] emphasise the existence of spatial auto-correlation in geographical data; however, cancer counts may be expected to be independent *conditionally* on different underlying risks, which admittedly may themselves exhibit spatial relationship. If the factors that account for these variations in risk are fully accounted for, the counts themselves should be independent provided that individual disease episodes arise independently of one another. The effect of assuming independence when these factors have not all been accounted for is another matter, but for many purposes it seems to us to be unlikely to be of great importance. In the case of the childhood L & NHL data, we have argued that the heterogeneity of risk is not large, particularly after fitting the social factors at our disposal, though admittedly we have not fully accounted for all apparent variation. The possibilities for improving on our control of confounding are limited. The Townsend index has a correlation with adult disease among the highest of the indices in common use; others, such as the Carstairs index, for example, use social class variables that are not easily available on a basis that is comparable between 1971 and 1981 censuses.

The question of the goodness of fit of a Poisson regression model when expectations are small is not straightforward. Some theoretical results are available[9] but they lead to difficult mathematics and are not easily implemented within standard packages. It should be remembered, though, that the ability to examine the goodness of fit of the model is not an option with standardisation, so the difficulties do not argue against modelling for controlling for confounding in the way we have outlined.

We thank Nuclear Electric plc for their financial support during part of the work described in this report, Dr Gerald Draper of the Childhood Cancer Research Group for his useful comments on an earlier draft of the paper and Professor Peter Diggle for a helpful discussion on the analysis. We also thank the Office of Population Censuses and Surveys, the Information and Statistics Division of the Common Services Agency of the Scottish Health Service, the Registrar General for Scotland, regional cancer registries, and the UK Children's Cancer Study Group for providing notifications of childhood cancer cases. The Childhood Cancer Research Group is supported by the Department of Health and the Scottish Home and Health Department.

1 Bithell JF, Dutton SJ, Draper GJ, Neary NM. Distribution of childhood leukaemias and non-Hodgkin's lymphomas near nuclear installations in England and Wales. *BMJ* 1994; **309**:501–5.
2 Office of Population Censuses and Surveys. *Changes in small areas 1971/81: Census Tracts/Parishes and the Change Files. Census 1981 User Guide 79.* London: OPCS, 1985.
3 Aitken M, Anderson D, Francis B, Hinde J. *Statistical modelling in GLIM.* Oxford: Clarendon Press, 1989.
4 Townsend P, Beattie A, Phillimore P. *Health and deprivation: inequality and the north.* London: Croom Helm, 1988.
5 Morris R, Carstairs V. Which deprivation? A comparison of selected deprivation indices. *J Public Health Med* 1992;**13**: 318–26.
6 Rodrigues L, Hills M, McGale P, Elliott P. Socio-economic factors in relation to childhood leukaemia and non-Hodgkin lymphomas: an analysis based on small area statistics for census tracts. In: Draper G, ed. *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966–83: Studies on Medical and Population Subjects, 53.* London: HMSO, 1991.
7 Diggle PJ. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *J R Stat Soc A* 1990;**153**:349–62.
8 Lawson AB. Using spatial Gaussian priors to model heterogeneity in environmental epidemiology. *The Statistician* 1994;**43**:69–76.
9 McCullagh P. The conditional distribution of goodness-of-fit statistics for discrete data. *J Am Stat Assoc* 1986;**81**: 104–7.

Open discussion

DIGGLE – Until we have resolved these problems, researchers with a good enough computing environment can always simulate the model, simulate the fitting process, and achieve an empirical distribution of the residual deviance. As you say, however, that may be a long way from the nominal χ^2 distribution.

BITHELL – But received wisdom is that you should be doing that simulation conditional on the observed values of the parameters you have estimated, and that is not easy.

DIGGLE – Yes, but I would include the variability of the parameter estimation in that simulation process. This is not an ideal solution, I know – it might be inefficient, but it would at least be valid.

BITHELL – I think that is a very good model but it is not actually received wisdom. In a situation like this it is a very complicated calculation.

OPENSHAW – It seems to me that the problem here is that you are studying fixed geography like electoral wards and then trying to cover bias by clever statistics. There is a simple solution and that is to re-engineer the geography to avoid the statistical problem – you reaggregate to areas that are somewhat similar to each other and above a certain minimum expectation or size. This avoids many of the problems involved with analysing spatial data, including the problems you describe.

BITHELL – I do not want to make the areas any bigger though, because I am trying to look at a high resolution areal effect.

OPENSHAW – But the problem is that there is nothing special about a ward from a geographical point of view. Some wards are small, some are large.

BITHELL – So you would construct artificial units rather as Black did to have similar expectations.[1]

OPENSHAW – Yes, and also to have some similarity in their characteristics: and that would reduce the confounding factors.

BITHELL – But even so you are going to end up with expectations in a study like this that are well under unity, and we do not wish to go to a bigger unit because we actually want to see what is happening on a small geographical scale. As long as expectations are less than unity there is still a problem with the deviance.

OPENSHAW – I have results that have an expectation greater than unity and the problem then disappears. Wards, as geographical objects, vary tremendously in size already. What is probably needed is to group the small zones to make them bigger so that they are somewhat more similar to the big zones already in the data set. So you are actually not losing very much and you probably are gaining a lot by simplifying.

ELLIOTT – You did a lot of regression, worried about the goodness of fit, and found quite a number of advantages over the stratification method, which I agree with in your case. May I ask you more generally about the SAHSU problem? We may have many different disease outcomes in any one study, possibly many different point sources, some of the disease outcomes may be rare and others common. You mentioned the problems of technology, both hardware and software, and told us that *GLIM* was able to cope. Would we be able to cope with a regression approach across the range?

BITHELL – I do not see why not. Specific to our problems were the more data related aspects, such as how strong an association we had and how big the expectations were. Unless you are looking at two orders of magnitude I can not see a problem with more data units. At ward level there is a maximum of 10 000 units and at enumeration district about 10 times that many. I do not know whether *GLIM* could handle enumeration districts for the whole of Britain but in any case you have regionalised your analysis to some extent. I do not think there is a technical difficulty in doing it for your work.

1 Committee on Medical Aspects of Radiation in the Environment. 2nd Report. *Investigation of the possible increased incidence of childhood cancer in young persons near the Dounreay nuclear establishment, Caithness, Scotland.* London: HMSO, 1988.
2 Openshaw S, Rao L. Algorithms for re-engineering 1991 census geography. *Environment and Planning A* 1995;27: 425–46.