ORIGINAL ARTICLE

Poisson regression analysis of ungrouped data

D Loomis, D B Richardson, L Elliott

Occup Environ Med 2005;**62**:325–329. doi: 10.1136/oem.2004.017459

See end of article for authors' affiliations

Correspondence to: Prof. D Loomis, Dept of Epidemiology, CB-7435 UNC-CH, Chapel Hill, NC 27599-7435, USA; Dana.Loomis@unc.edu

Accepted 13 December 2004

Background: Poisson regression is routinely used for analysis of epidemiological data from studies of large occupational cohorts. It is typically implemented as a grouped method of data analysis in which all exposure and covariate information is categorised and person-time and events are tabulated.
Aims: To describe an alternative approach to Poisson regression analysis using single units of person-time without grouping.
Methods: Data for simulated and empirical cohorts were analysed by Poisson regression. In analyses of simulated data, effect estimates derived via Poisson regression without grouping were compared to those obtained under proportional hazards regression. Analyses of empirical data for a cohort of 138 900 electrical workers were used to illustrate how the ungrouped approach may be applied in analyses of actual occupational cohorts.
Results: Using simulated data, Poisson regression analyses of ungrouped person-time data yield results equivalent to those obtained via proportional hazards regression: the results of both methods gave unbiased estimates of the ''true'' association specified for the simulation. Analyses of empirical data confirm that grouped and ungrouped analyses provide identical results when the same models are specified. However, bias may arise when exposure-response trends are estimated via Poisson regression analyses in which exposure scores, such as category means or midpoints, are assigned to grouped data.
Conclusions: Poisson regression analysis of ungrouped person-time data is a useful tool that can avoid bias associated with categorising exposure data and assigning exposure scores, and facilitate direct assessment of the consequences of exposure categorisation and score assignment on regression results.

Poisson regression is a method of modelling disease rates as a function of covariate levels that is often applied in the analysis of data from occupational cohort studies.[1] Analyses are typically conducted using grouped input data in the form of a tabulation of person-time and events in which all predictor variables are categorised.[2–5] Although categorisation of an exposure indicator is sometimes criticised, it remains useful, and in some circumstances even preferable to analyses of exposure data in continuous form.[1 6 7] However, for the purpose of estimating quantitative exposure-response relations, categorisation of exposure data that were originally measured on a continuous scale often leads to loss of power and questions about the sensitivity of study findings to decisions about exposure categorisation and score assignment.[7–10]

One way to address concerns about the consequences of exposure categorisation is to utilise a regression method, such as Cox proportional hazards regression, that accommodates continuous data.[2 11] However, proportional hazards regression methods can be extremely intensive computationally for analyses of large occupational cohorts. This is particularly true for analyses involving interactions and time dependent variables; in such cases Cox regression models may fail to converge.[11]

In this paper we describe how Poisson regression analyses using single units of person-time, rather than the standard grouped person-time approach, may be used to directly evaluate these concerns. This ungrouped approach avoids the need to categorise variables originally measured on a continuous scale and facilitates examining the influence on regression results of exposure categorisation and score assignment. The researcher can use the same regression model and methods applied for analyses of grouped data, but without categorisation of predictor variables. In addition, Poisson regression allows the rate ratio, a fundamental epidemiological indicator, to be estimated directly from the data. We illustrate the ungrouped Poisson regression method and its application through simulations and analyses of empirical data from a large occupational cohort.

METHODS

Poisson regression

The assumption of a classical Poisson regression model is that the number of events in a particular unit of time follows the Poisson distribution with a mean $n\lambda$, where for observation i the rate λ_i is related to a vector of independent explanatory variables, $\mathbf{X}_i$, by

$$\log(\lambda_i) = \log(n_i) + \mathbf{X}_i \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters to be estimated and n_i represents the time at risk and is equivalent to the rate denominator. The quantity $\log(n_i)$ is often referred to as the ''offset'' of the model. With this model, a cohort is typically cross-classified by levels of exposure and other predictor variables, $\mathbf{X}_i$, and the time at risk is calculated for each of the resulting combinations of $\mathbf{X}$.

Ungrouped input structure for person-time data

In order to conduct Poisson regression analysis of ungrouped person-time data, an analytical data set is constructed in which there is a unique observation for each unit of person-time at risk. An example of the analytical data for one subject from a simulated cohort (described in detail below) is shown in table 1. One subject contributes multiple observations, and the total number of observations is equal the total person-years of follow up. As indicated in column 2 of the table, a binary indicator of case status is associated with each observation. The indicator is assigned a value of ''0'' for each observation (each unit of person-time at risk) until the date of last observation; at that point, a value of ''1'' is assigned to cases (and a value of ''0'' is assigned to non-cases). As indicated in column 3, each observation represents one unit

Main messages

- Poisson regression models can be fit to ungrouped person-time data, as well as to input data in the traditional, tabular form.
- With ungrouped input data, exposure need not be categorised, but can instead be expressed as a continuous, quantitative variable.
- Ungrouped Poisson and Cox regression models give equivalent results, but Poisson regression directly estimates rate ratios and may have advantages in computational efficiency.
- The ungrouped approach can avoid bias associated with exposure categorisation.
- Poisson regression models based on grouped and ungrouped data provide identical estimates of exposure-disease association and precision when the models are equally specified.

Policy implications

- Ungrouped Poisson regression may be a preferred approach for risk assessment.

of person-time (in this example, one year of follow up). Therefore, all observations contribute equal weight and the offset term need not be specified when fitting the Poisson regression model to ungrouped data.

As shown in columns 4–5 of table 1, each observation can be associated with independent predictor variables that are measured on a continuous scale. Column 4 illustrates how each unit of person-time is associated with an attained age (measured on a continuous scale). Column 5 shows how each observation is also associated with a cumulative exposure level.

Simulation

Hypothetical data were generated for 100 cohorts, each with 25 000 workers. At the start of follow up, each simulated worker was assigned an age-at-entry into the cohort, maximum lengths of follow up and employment, and an exposure rate equal to the amount of exposure accumulated in one year (table 2). The distribution of age-at-entry and lengths of follow up and employment are similar to those observed in a study of nuclear industry workers.[12] The median age at entry is 25 years, while the 90th centile for age-at-entry is 41 years. The median lengths of employment and follow up are 17 years and 35 years, respectively. For each person-year of observation contributed by a simulated subject, disease status was determined by calculating the probability of disease under the model:

$$\log (p) = \delta_0 + \delta_1 \text{age} + \phi x,$$

where δ_0 and δ_1 are parameters for a Weibull model centred at age 55 years that defines the age specific probability of disease in the absence of exposure, where x and *age* are time dependent indicators representing cumulative exposure and the natural logarithm of attained age, respectively, and ϕ is the effect of exposure on the probability of disease. We allowed for censoring of observations because of death due to causes other than the one under investigation by calculating, for each person-year, the age specific probability of censoring via the model:

$$\log (c) = \eta_0 + \eta_1 (\text{age})$$

where c is the probability of censoring, and η_0 and η_1 are parameters of a Weibull model that defines the age specific risk of the death due to causes other than the one under investigation. Values of δ_0, δ_1, η_0, η_1, and ϕ are given in table 2. Further details of the simulation methods used here (and an example of the SAS code used to generate simulated cohort data) are given in a previous publication.[13]

Empirical data

Data for empirical analyses were obtained from a retrospective study of mortality among a cohort of 138 905 male electrical workers in the United States. Details of the study, which was originally designed to examine the risk of leukaemia and brain cancer in relation to exposure to magnetic fields, have been presented elsewhere.[14] Briefly, the men were employed for at least six months at any of five electric power companies between 1950 and 1986 and were followed through 1988, yielding 20 733 deaths. Exposure to 60 Hz magnetic fields was estimated by linking individual work histories with quantitative data derived from 2842 full-shift personal magnetic field measurements.[15 16] The large size and unusually complete follow up (97%) of this cohort make it particularly useful for methodological research. For the purpose of the current analysis, we considered the association of brain cancer with exposure to magnetic fields, estimated as unlagged cumulative exposure in micro Tesla-years (μT-y). Note that the risk estimates obtained here are not necessarily equal to those published previously because of differences in parameterisation and model specification.

Data analysis

Poisson regression models were fit to the simulated and empirical data. The simulated data were entered in ungrouped form, as described above, and no offset term was specified. Age and exposure were the only explanatory variables in analyses of simulated data.

Poisson regression analyses of empirical data from the electrical workers cohort were conducted with the input data entered both in ungrouped form and in the classical, tabular form. When the tabular input form was used, all of the predictor variables, including exposure, were categorised and an offset term was included in the model. Quantitative exposure scores for categorical analyses were assigned by dividing the data at deciles of the exposure distribution among all person-years or among person-years of brain cancer cases only, and then selecting the mean exposure level of each category to represent the exposure of all events and person-time at risk in that category. We also considered scores based on category midpoints. However, we showed in a previous paper[10] that midpoint exposure scores tend to increase the bias resulting from categorisation, so in the interest of brevity those data are not shown here. When the ungrouped form of input was used, exposure was also entered as a continuous variable. Models fit to the cohort data were adjusted for age and calendar time, which were categorised in 10 year increments when grouped and ungrouped models were compared. Race (in two categories) was considered as an additional predictor in some analyses to approximate the complexity of typical occupational cohort analyses.

Proportional hazards regression was also used to derive estimates of cumulative exposure-mortality trends using the simulated and empirical data. Attained age was specified as the timescale to obtain relative risk estimates. Cumulative exposure was treated as continuous variable and, in analyses of empirical cohort data, calendar time and race were

Table 1 Simulated cohort data showing the ungrouped data structure for one subject who began exposure at age 42 and developed the disease of interest at age 64

Observation	Case status	Person-time (years)	Age at risk	Cumulative exposure
1	0	1	42	0.31391
2	0	1	43	0.62782
3	0	1	44	0.94173
4	0	1	45	1.25564
5	0	1	46	1.56955
6	0	1	47	1.88346
7	0	1	48	2.19737
8	0	1	49	2.51128
9	0	1	50	2.82519
10	0	1	51	3.13910
11	0	1	52	3.45301
12	0	1	53	3.76692
13	0	1	54	4.08083
14	0	1	55	4.39474
15	0	1	56	4.70865
16	0	1	57	5.02256
17	0	1	58	5.33647
18	0	1	59	5.65038
19	0	1	60	5.96429
20	0	1	61	6.27820
21	0	1	62	6.59211
22	0	1	63	6.90602
23	1	1	64	7.21993

included as additional explanatory variables to match the Poisson regression models.

The SAS system (SAS Institute, Cary, North Carolina, USA) was used to generate the simulated cohorts, compute person-time at risk, and fit the regression models.

RESULTS

Estimates of the exposure-disease association in simulated data were obtained using proportional hazards regression and Poisson regression analyses of ungrouped person-time data. Analyses by both methods yielded quantitatively similar results, as indicated in fig 1 by the alignment of the estimates along a line of equality. The average estimate via each method was 0.40 (diamond in fig 1), equal to the true magnitude of association specified in the simulation.

Parallel analyses of the occupational cohort data using proportional hazards regression and Poisson regression with ungrouped input data also yielded identical estimates of the exposure-disease association and its standard error (table 3).

To evaluate the effect of exposure categorisation and score assignment, Poisson regression models were also fit to grouped person-time data from the occupational cohort. In the grouped data, cumulative exposure was represented by exposure scores based on mean values for categories defined by deciles of the exposure distribution among all person-years or deciles of the exposure distribution among cases. Estimates of the association based on categorised exposure were different from those obtained with a continuous exposure variable and ungrouped Poisson regression or proportional hazards regression (table 3), suggesting that results obtained with the categorical approach are biased. The apparent bias was reduced by using the distribution of exposure among cases, rather than among all person-years, as the basis for categorisation (table 3).

We also fit identically specified Poisson regression models to the empirical cohort data in tabular and ungrouped form. This required that exposure and all covariates be categorised, because the tabular input form cannot accommodate continuous variables. These models yielded identical estimated RRs and 95% confidence intervals for both forms of input (data not shown).

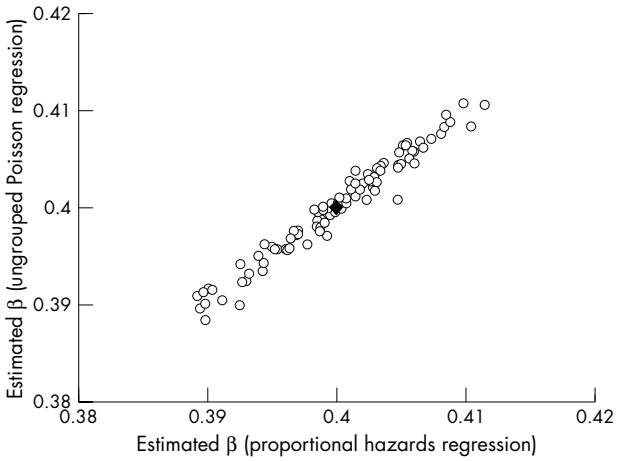

Figure 1 Estimates of dose-response trends derived via Poisson regression of ungrouped person-time data and proportional hazards regression.

Table 2 Conditions specified for simulation

	Condition
Number of iterations of simulation	100
Number of persons in study cohort	25000
Age at entry (in years)	$18 + 10(Exp(1))$
Length of follow up (in years)	$40 - 5(Exp(1))$
Length of employment (in years)	$25Exp(1)$
Disease rate in the absence of exposure	$\delta_0 = -6.3$ $\delta_1 = 5.7$
Probability of censoring from other causes	$\eta_0 = -5.0$ $\eta_1 = 5.1$
Exposure effect	$\phi = 0.400$
Occupational exposure rate	$Exp(1)$

The exponential distribution, with mean and variance equal to 1, is denoted $Exp(1)$.

Table 3 Comparison of estimated regression beta coefficients* and standard errors for brain cancer and cumulative magnetic field exposure in the electrical worker cohort

	Ungrouped Poisson	Proportional hazards	Grouped Poisson	
Model	β (SE)	β (SE)	β† (SE)	β‡ (SE)
Model 1: age, exposure	0.0842 (0.0375)	0.0842 (0.0375)	0.1483 (0.0503)	0.1123 (0.0392)
Model 2: age, calendar time, race, exposure	0.0910 (0.0380)	0.0910 (0.0380)	0.1479 (0.0508)	0.1183 (0.0396)

From proportional hazards and ungrouped Poisson regression with continuous exposure variables and grouped Poisson regression with quantitative exposure scores assigned by two methods, as shown.
*Beta coefficients represent the estimated change in the log rate or hazard or brain cancer per μT-year exposure to 60 Hz magnetic fields.
†Exposure categories defined by deciles of the population distribution of exposure. Scores assigned to exposure categories based on mean values for all person-time accrued in the category.
‡Exposure categories defined by deciles of the distribution of exposure among cases. Scores assigned to exposure categories based on mean values for all person-time accrued in the category.

DISCUSSION

We propose that Poisson regression analysis of ungrouped person-time data can be used to estimate quantitative exposure-response relations and address concerns about potential bias resulting from the definition of exposure variables. This approach allows Poisson regression models to be applied to occupational cohort data with exposure estimates entered as a continuous variable. As a result, it facilitates the comparison of alternative forms of exposure-response, ranging from categorisation to parametrically and non-parametrically smoothed curves.[7 17] It also permits the investigator to use the same regression method, and in fact the identical regression models, to examine the same data, in an ungrouped format, that are analysed in classical Poisson regression in a tabular form. To our knowledge, the ungrouped approach to Poisson regression has not been described previously, although it may have been applied in a recent analysis of occupational cohort data.[18]

It has been asserted that the Poisson regression method is equivalent to the risk set approach of Cox proportional hazards regression under the situation in which each cell of the cross-classification of person-time and events includes a single event.[1 11] This assertion is correct for the situation in which estimates of association are derived solely using categorical variables. However, if scores are assigned to exposure categories in order to estimate dose-response trends, then these approaches are not equivalent. Even if each cell of the person-time table includes a single death it may include multiple person-years at risk. The score assigned to that cell will not necessarily provide an unbiased estimate of the true exposure for the person-time and the decedent in each cell. In contrast, the ungrouped approach we describe, in which each unit of person-time and each event is associated with its measured exposure, will in fact converge to the risk set approach to analyses of continuous data as units of person-time become increasingly small. We have shown through simulation and analyses of empirical occupational cohort data that the two methods yield equivalent results when applied to the same data.

When exposure and disease occurrence are quantitatively related, categorisation of a continuous exposure variable may produce differential misclassification and bias estimates of association in a positive or negative direction.[8 19–21] In an earlier paper, we illustrated the operation of this bias in exposure-response analyses in which categories of exposure are represented by assigned scores.[10] Exposure scores derived from the category midpoints produce negative bias, while scores based on category means, as in the examples in this paper, bias associations in a positive direction. The bias is likely to be small if exposure categories are narrowly defined and scores are assigned based on person-time weighted mean values.[10] Nonetheless, concerns about the consequences of exposure categorisation and score assignment arise frequently in quantitative risk assessment; the literature includes many examples of researchers evaluating the sensitivity of study results to categorisation by varying boundaries and rules about score assignment.[22–26] Such approaches are time consuming and ultimately never conclusive. An approach that does not require categorisation would clearly be useful.

The approach to Poisson regression analysis of ungrouped person-time data we describe can be used to test the sensitivity of the results to the investigator's decisions about exposure categorisation and score assignment. There are distinct advantages in retaining the Poisson regression approach for this purpose. An alternative would be to use a different method, such as Cox proportional hazards or conditional logistic regression, to evaluate questions about the impact on risk estimates of categorising exposure data. However, this would entail a different regression model with different assumptions. Furthermore, in our experience, while contemporary computers have substantially reduced the obstacles to fitting Cox regression models for analyses of data from studies of large occupational cohorts, computational obstacles remain when attempting to fit models that involve interactions between a time dependent variable and the timescale specified for the Cox model (for example, an interaction between a baseline timescale of attained age and a time dependent indicator of active employment status). Again, in such cases, the approach to Poisson regression analysis of ungrouped person-time data offers a simple, direct way to evaluate the sensitivity of the results to decisions about exposure categorisation and score assignment.

The ungrouped method that we illustrate sacrifices much of the computational efficiency obtained by grouped data analyses of cohort data. Poisson regression analyses in which ungrouped data are generated with a unique observation for each person-day at risk are typically the most refined classification of study data. In many cases, computational efficiency can be gained by a less refined categorisation of person-time (for example, person-years). As illustrated in this paper, however, given the advances in the processing speed of personal computers it is no longer necessary to limit analyses to grouped data approaches to Poisson regression.

ACKNOWLEDGEMENTS

We thank John Bailer, David Kriebel, Steve Marshall, and Bob Park for constructive comments on earlier versions of this paper.

Authors' affiliations

D Loomis, D B Richardson, Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA
L Elliott, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

Competing interests: none

REFERENCES

1 **Checkoway H**, Pearce N, Kriebel D. *Research methods in occupational epidemiology*, 2nd edn. New York: Oxford University Press, 2004.
2 **Breslow NE**, Day NE. *Statistical methods in cancer research: the design and analysis of cohort studies*. Lyon: International Agency for Research on Cancer, 1987.
3 **Checkoway H**, Pearce N, Dement JM. Design and conduct of occupational epidemiology studies: I. Analysis of cohort data. *Am J Ind Med* 1989;**15**:375–94.
4 **Frome EL**, Checkoway H. Epidemiologic programs for computers and calculators. Use of Poisson regression models in estimating incidence rates and ratios. *Am J Epidemiol* 1985;**121**:309–23.
5 **Krewski D**, Cardis E, Zeise L, *et al*. Empirical approaches to risk estimation and prediction. In: Moolgavkar SH, Krewski D, Zeise L, Cardis E, Moeller H, eds. *Quantitative estimation and prediction of human cancer risks*. Lyon: International Agency for Research on Cancer, 1999:31–178.
6 **Greenland S**. Dose-response and trend analyses in epidemiology: alternatives to categorical analysis. *Epidemiology* 1995;**6**:356–65.
7 **Steenland K**, Deddens JA. A practical guide to dose-response analyses and risk assessment in occupational epidemiology. *Epidemiology* 2004;**15**:63–70.
8 **Flegal KM**, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. *Am J Epidemiol* 1991;**134**:1233–44.
9 **Wartenberg D**, Northridge M. Defining exposure in case-control studies: a new approach. *Am J Epidemiol* 1991;**133**:1058–71.
10 **Richardson DB**, Loomis D. The impact of exposure categorisation for grouped analyses of cohort data. *Occup Environ Med* 2004;**61**:930–5.
11 **Pearce N**, Checkoway H, Dement J. Exponential models for analyses of time-related factors, illustrated with asbestos textile worker mortality data. *J Occup Med* 1988;**30**:517–22.
12 **Richardson DB**, Wing S. Greater sensitivity to ionizing radiation at older age: follow-up of workers at Oak Ridge National Laboratory through 1990. *Int J Epidemiol* 1999;**28**:428–36.
13 **Richardson DB**. Power calculations for survival analyses via Monte Carlo estimation. *Am J Ind Med* 2003;**44**:532–9.
14 **Savitz DA**, Loomis DP. Magnetic field exposure in relation to leukemia and brain cancer mortality among electric utility workers. *Am J Epidemiol* 1995;**141**:123–34.
15 **Loomis DP**, Peipins LA, Browning SR, *et al*. Classification and organization of work history data in industry-wide studies: an application to the electric power industry. *Am J Ind Med* 1994;**26**:413–25.
16 **Kromhout H**, Loomis DP, Mihlan GJ, *et al*. Assessment and grouping of occupational magnetic field exposure in five electric utility companies. *Scand J Work Environ Health* 1995;**21**:43–50.
17 **Thurston SW**, Eisen EA, Schwartz J. Smoothing in survival models: an application to workers exposed to metalworking fluids. *Epidemiology* 2002;**13**:685–92.
18 **McDonald JC**, Harris J, Armstrong B. Mortality in a cohort of vermiculite miners exposed to fibrous amphibolite in Libby, Montana. *Occup Environ Med* 2004;**61**:363–6.
19 **Dosemeci M**, Wacholder S, Lubin JH. Does nondifferential misclassification always bias a rue effect toward the null value? *Am J Epidemiol* 1990;**132**:746–8.
20 **Brenner H**, Loomis D. Varied forms of bias due to nondifferential error in measuring exposure. *Epidemiology* 1994;**5**:510–17.
21 **Steenland K**, Deddens JA, Zhao S. Biases in estimating the effect of cumulative exposure in log-linear models when estimated exposure levels are assigned. *Scand J Work Environ Health* 2000;**26**:37–43.
22 **Wing S**, Shy CM, Wood JL, *et al*. Mortality among workers at Oak Ridge National Laboratory. Evidence of radiation effects in follow-up through 1984. *JAMA* 1991;**265**:1397–402.
23 **Gilbert ES**, Cragle DL, Wiggs LD. Updated analyses of combined mortality data for workers at the Hanford site, Oak Ridge National Laboratory, and Rocky Flats weapons plant. *Radiat Res* 1993;**136**:408–21.
24 **Lea CS**, Buffler PS, Merrill DW, *et al*. Reassessment of cancer mortality among employees at Oak Ridge National Laboratory with follow-up through 1984: a comparison with results of previously published studies. *Technology* 2000;**7**:303–16.
25 **Loomis D**, Kromhout H, Kleckner RC, *et al*. Effects of analytical treatment of exposure data on associations of cancer and occupational magnetic field exposure. *Am J Ind Med* 1998;**34**:49–56.
26 **Kheifets LI**, Gilbert ES, Sussman SS, *et al*. Comparative analyses of the studies of magnetic fields and cancer in electric utility workers: studies from France, Canada and the United States. *Occup Environ Med* 1999;**56**:567–74.