Relative Risk Estimation for Occupational Groups in Epidemiologic Surveillance

E. L. Frome, Oak Ridge National Laboratory
Computer Science and Mathematics Division
C. H. Strader and H. G. Stockwell, Department of Energy
E.A. Dupree, Oak Ridge Institute for Science and Education
E. L. Frome, ORNL, P.O. Box 2008, Oak Ridge,TN 37831-6367

Key Words:epidemiology; Poisson regression; epidemiologic surveillance; relative risk; occupational cohorts; disease rates

From the 1999 PROCEEDINGS OF THE STATISTICS IN EPIDEMIOLOGY SECTION OF THE AMERICAN STATISTICAL ASSOCIATION

Abstract:

The Department of Energy's Office of Epidemiologic Studies (http://tis.eh.doe.gov/epi/) uses epidemiologic surveillance to monitor morbidity and assess the overall health of the work force. Annual reports summarize the results of epidemiologic surveillance at participating sites. This paper describes methods for the analysis of disease specific relative risk experienced by occupational groups. In this situation, there is no logical nonexposed (i.e., referent) group. Poisson regression is used to develop a method with the entire cohort as the referent group, and considers age and gender as potential confounding variables and/or effect modifiers. A log-linear regression function describes the effect of the explanatory variables on disease rates. An example is presented to illustrate the methods. Links to more detailed information and data that are available on the Internet are provided at http://www.epm.ornl.gov/~frome/.

1. Introduction

The mission of the U.S. Department of Energy's [1](DOE) epidemiologic surveillance program is to monitor morbidity, assess the overall health of the work force, and identify groups that may be at increased risk for occupation-related injury or illness. The program provides a focus for actions that may reduce or eliminate risk. Epidemiologic surveillance also provides a means for measuring the effectiveness of corrective actions. This is accomplished through the routine collection, analysis, and interpretation of selected morbidity, demographic, and occupational exposure data. Additional information is available under *Intramural Health Studies* at DOE's Epidemiologic Studies Internet site (http://tis-nt.eh.doe.gov/epi). The epidemiologic surveillance report web site (ESRWS) at http://www.epm.ornl.gov/~frome/ES presents a detailed description of an annual report.

This paper presents a method for the estimation of annual age-adjusted diagnostic rates and relative risks for occupational groups. A log-linear Poisson regression model is used to estimate parameters of interest, (e.g., relative risks for occupational groups) and to "adjust" for explanatory variables of secondary interest (age and sex) that may be confounders and/or effect modifiers.

1.1 Description of Data

Data are collected by coordinators at each DOE site and submitted to the Epidemiologic Surveillance Data Center (ESDC), located at the Oak Ridge Institute for Science and Education (ORISE). The ESDC is responsible for quality control, record linkage, and nosological coding of the diagnostic data according to the ICD-9-CM. The source data files obtained obtained annually from each facility are (1) a roster file, (2) a five day absence file, and (3) an OSHA-Recordable events file. The later two files contain *health events*. There is at least one disease or injury diagnosis associated with each health event. The event files and the roster are used to generate stratified tables that serve as the starting point for Poisson regression.

2. Statistical Analyses

The theoretical and computational basis for using regression methods to obtain maximum likelihood estimates of parameters in regression models for Poisson distributed data—i.e., Poisson regression (PR)—was given by Frome *et al* [11]. For application of PR to event rate

[1]Abbreviations used: AIC; Akakie Information Criterion; df, degrees of freedom; DOE, Department of Energy; ESRWS, epidemiologic surveillance report web site; ESDC, Epidemiologic Surveillance Data Center; GLM, generalized linear model; LRR, log relative risk; ORISE, Oak Ridge Institute for Science and Education; ORNL, Oak Ridge National Laboratory; pyrs, person-years; PR, Poisson regression; RR, relative risk; SE, standard error.

analysis see Frome [7] and for discussion of epidemiologic applications see [8, 12, 10]. Clayton and Hill [4] present an excellent discussion of PR and its advantages over stratified analysis. Breslow and Day [2] provide a good intermediate-level treatment of PR in epidemiologic studies, and McCullagh and Nelder [14] provide a more general mathematical treatment of generalized linear models (GLM).

2.1 Poisson Regression Model For Event Rate Data

Let y_{ijk} denote the number of diagnoses in the ith age group, j^{th} sex , and k^{th} occupational category for a given calendar year. The y_{ijk} are treated as observed values of Poisson variates with expectation

$$E(y_{ijk}) = n_{ijk}\lambda_{ijk} = n_{ijk}exp(X\beta), \qquad (1)$$

where $i = 1,\ldots,I$; $j = 1,2$; and $k = 1,\ldots,K$. In Eq. (1), λ_{ijk} is an expected rate (defined by a regression function), X is a row vector of explanatory variables, β is the vector of regression parameters (see below), and n_{ijk} denotes the person-years (in 1000s) so that rates are per thousand person-years (pyrs). The value of each row X of the model matrix is determined by the factor levels for that observation and the constraints that are used. Eq. (1) is generally referred to as a multiplicative or log-linear regression function.

2.2 All Explanatory Variables as Factors

A standard notation in the statistical literature for a log-linear PR model with explanatory variables age, sex, and occupational group considered as factors is

$$log(\lambda_{ijk}) = \mu + \alpha_i + \theta_j + \delta_k. \qquad (2)$$

This *main effects model* includes all three explanatory variables, but has no *interaction* terms. The explanatory variable of primary interest (i.e., the *exposure* variable) is occupational group, and age and sex are viewed as confounders and/or effect modifiers. In Eq. (2), δ_ks represent occupational group effects, α_is age effects, θ_js sex effects, and μ is the "overall rate"—all on on a logarithmic scale. For practical reasons [4] it is more convenient to write Eq. 2 as

$$log(rate) = r + ag + s + og, \qquad (3)$$

where ag is the age factor with I levels, s denotes sex with two levels, og is occupational group with K levels, and r corresponds to μ. This notation corresponds to that used in statistical programs that fit generalized linear models such as GLIM [6] and Splus [3, 16]. For example, in computer output the value of $a(i)$ corresponds to the estimate of α_i in Eq. (2) notation. The parameters in Eq.(2) are not all estimable, and constraints are imposed by the coding used to define the rows—X in Eq.(1)—of the model matrix.

2.2.1 Conventional Constraints Used in GLMs

The usual convention used to impose constraints when fitting log-linear models—see e.g. [2, 4, 6]—is to take the first (or some other appropriate) level of each factor as the *baseline* rate (i.e., set $a(1) = s(1) = og(1) = 0$). In this approach, r represents the log of the baseline rate when all factors are at their first (referent) level. If the *exposure* variable has a "natural" referent level (i.e., level one corresponds to the *nonexposed group*), this approach is reasonable, and $og(2)$, ...,$og(K)$ are log relative risk (LRR) estimates with group 1 as the referent group. The referent level for age and sex is usually unimportant and can be changed to any convenient value.

In the epidemiologic surveillance situation there is no particular occupational group that is appropriate as the *nonexposed group* for each disease or injury category. One way to deal with this is to define an indicator variable for the kth occupational group, such as $xg = 1$ if $og = k$ and xg =0 otherwise. Next, fit the model $log(rate) = ag * s + xg$, where $ag * s$ is shorthand for $r + ag + s + ag : s$. The parameter estimate corresponding to xg is the LRR for the kth occupational group (i.e., the "exposed") relative to all other workers (the "nonexposed") adjusted for age and sex. This is equivalent to the method described by Greenland and Sanders [13]. They also discuss the closed-form stratified Mantel-Haenszel rate ratio estimator proposed by Rothman and Boice [15]. The indicator variable approach leads to relative risk estimates that are difficult to interpret since the referent group changes for each value of k. This stratified analysis also precludes the possibility of looking at interactions, e.g. between occupational group and sex, that may be of practical interest.

2.2.2 The Sum To Zero Constraints

When there is no logical control group, the *effects sum to zero* restrictions can be used. Specifically, assume that the $a(i)$s, $s(j)$s, and $og(k)$s each sum to zero. These are the classical assumptions used in the analysis of variance for linear models—see e.g. Section 9.7, Draper and Smith [5]. In this situation:

- r is the log of the *overall disease rate* (1 parameter),
- $ag(i)$ is the ith age group effect (I parameters) ,
- $s(j)$ is the effect of jth sex (2 parameters), and
- $og(k)$ is the kth occupational group effect (K parameters).

The parameters $og(k)$ represent the amount by which the kth occupational group differs from the overall rate on the log scale. Exp(r) is an estimate of *overall annual disease rate*, and for a given diagnostic category, exp[$og(k)$] is the age and gender adjusted relative risk for the kth occupational group with all employees at the site as the comparison group. Estimates of the LRR parameters (adjusted for age and gender) are obtained by fitting the main effects model in Eq.(2). An alternative approach is to use a stratified (on age and gender) analysis. This is equivalent to fitting the main effects model with an age-gender interaction term added to the the model:

$$log(rate) = r + ag + s + ag.s + og. \qquad (4)$$

2.2.3 Representation of Relative Risk in Log Percent (L%) units

The parameter $og(k)$ can be viewed as the difference between the kth occupational group and the overall rate on a logarithmic scale (i.e., the log difference). Tornquist, Vartia and Vartia [17] have discussed the use of the log difference as a measure or relative difference in an economic context, and proposed the convention of multiplying the values by 100 and calling this log percent denoted by the symbol L%. They show that log difference is the only symmetric, additive, and normed indicator of relative change. Further discussion and examples of the use of this approach to describing relative risks in log percent units in an epidemiologic context are given by Frome *et al* [9, 10] (see www.epm.ornl.gov/~frome/ORMS/repht/node22.html)

3. Results

The data in the example was obtained from the ESDC and was used to calculate the relative risk results in the 1993 Rocky Flats Annual Report using the indicator variable method. Tables stratified by age, sex, and occupational group for several diagnostic categories are available at ESRWS (see *Rocky Flats Annual Report for 1993*) , along with a description of the Rocky Flats Plant and other material included in the annual DOE/EH-62 epidemiologic surveillance report. Theses tables are based on health events which are defined as a period of five or more consecutive work-days away from work due to an illness or injury. There were 1,113 health events and the total person-years for the workforce was 6,348 (50 person-years with an unknown occupational category were excluded from this analysis).

Data for diseases of the respiratory system are used to illustrate the method. The LRR estimates in L% units are given for several models in Table 1. The results on line 1 are unadjusted estimates of the occupational group

Table 1: Analysis of Deviance Table For Diseases of the Respiratory System. The Deviance for the Main Effects Model (MEM) is 69.84.

Model/Term [a]	**df** [b]	**Delta Dev**[c]	**Prob**[d]	**AIC**[e]
1 ag + s + og	49	0.0	0.0	91.84
2 remove ag	4	+4.991	0.2882	88.83
3 remove s	1	+80.272	0.0000	170.11
4 remove og	5	+20.046	0.0012	101.89
5 add ag:s	4	-4.396	0.3551	95.44
6 add ag:og	20	-29.702	0.0748	102.14
7 add og:s	5	-11.952	0.0355	89.89

[a]The first row is for the main effects model.
[b]Degrees of freedom for term added.
[c]Change in Deviance from MEM for term in column 1.
[d]Probability for term added.
[e]Akaike Information Criteria= Deviance + 2df.

LRR parameters. From Row 2 of Table 1, the overall rate for the main effects model in Eq. (2) is exp(4.37)= 79/1000 pyrs. The age and gender adjusted LRR for the six occupational groups are given on line 2 of Table 1, and are shown graphically with 95% confidence limits in Figure 1. For example, for Technical workers og(3)= 0.2309; this is written as 23.09L% (read as 23.09 log percent). The conventional relative risk estimate is exp(0.2309)= 1.26 and the "adjusted rate" for og(3) is 1.26*79 = 100/1000 pyrs. If the first group is viewed as the referent group, then the LRR for the kth group is easily obtained by subtraction. For example (see Table 1 Row 2) if Technicians are compared to Administrative workers the LRR is 23.09 - (-21.41) = 45.68L%, corresponding to a RR of 1.58. Estimates of the log relative risk parameters for og based on the stratified analysis—see Eq. (4)—are on line 3b and are in close agreement with the results for the main effects model.

The analysis of deviance (AOD) table for the example is given in Table 2. The main effects model is used as the reference model for several likelihood ratio tests. The effect of removing factors is shown on lines 2-4, and the effect of adding interaction terms is shown on lines 5-7. The deviance for the main effects model is 69.84 with 49 df, providing some indication of lack-of-fit. From line 3 of Table 2, gender is seen to be the strongest explanatory variable. The value of the LRR estimate for females is s(1)= 47.2L% (SE= 5.19), and the corresponding rate (adjusted for age and og) is exp(0.472)*79= 126/1000 pyrs. Since, by definition, s(2)= -s(1), the adjusted rate for males is 49/1000 pyrs. In conventional terms, the respiratory disease health event rate is 2.6 times higher for females than for males.

Lines 5-7 in Table 2 show the effect of adding inter-

Table 2: Estimates of Log Relative Risk (L% Units) For Diseases of the Respirataory System By Occupational Groups

Model	r[a]	Admn[b]	Prof	Tech	Serv	CrML	Nucl
0) r alone	4.25	0	0	0	0	0	0
1) r+og	4.27	9.40	-32.36	27.36	44.18	-34.37	-14.21
2) r+og+ag+s	4.37	-21.41	-23.30	23.09	48.02	-14.09	-12.32
3a) r+ ag:s	4.29	0	0	0	0	0	0
3b)r+og+ag:s	4.37	-22.21	-22.71	23.47	48.04	-13.58	-13.02

[a] r = log Overall Disease Rate, e.g. exp(4.37)= 79 events per 1000 person years
[b] Example: For Model 2 the RR for the Admin Group is RR= exp(-.2141)= 0.807

action terms to the main effects model. The og:s term is large, indicating that a more detailed analysis may be of interest. The last column of Table 2 gives the value of the Akaike information criterion (AIC) for each model in Table 2. The AIC—see [1, 12, 14]—combines a measure of the discrepancy between the fitted values and the data (as measured by the deviance), and the simplicity of the model (as reflected by the number of parameters). The AIC is used to identify ”good models” and smaller is better. The smallest values for the AIC occur for the models that include occupational group and gender.

In summary, PR using the main effects model in Eq. (2) with the effects sum to zero restrictions provides an effective data analysis method for routine use in epidemiologic surveillance. For each disease category estimates of the overall adjusted rate and the LRR for each occupational group provide summary information for a given facility in a specific year. The AOD table is used to further evaluate alternative models that may be of interest and to determine the relative importance of each of the explanatory variables. An important advantage of PR is that it can be easily extended to combine data over time and DOE facilities when compatible data structures are available.

4. Acknowledgment and Disclaimer

This research was supported by the Offices of Epidemiologic Surveillance, Environment, Safety and Health, U. S. Department of Energy under contract DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp. The authors thank Drs. D. L. Cragle and D. M. Flanagan for reviewing this paper, the staff of ORISE’s Center for Epidemiologic Research for their help in data collection, and K. Miller for assistance in manuscript preparation.

This report was prepared as an account of work sponsored by an agency of the United States Government. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this work, or to allow others to do so for U. S. Government purposes. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express of implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

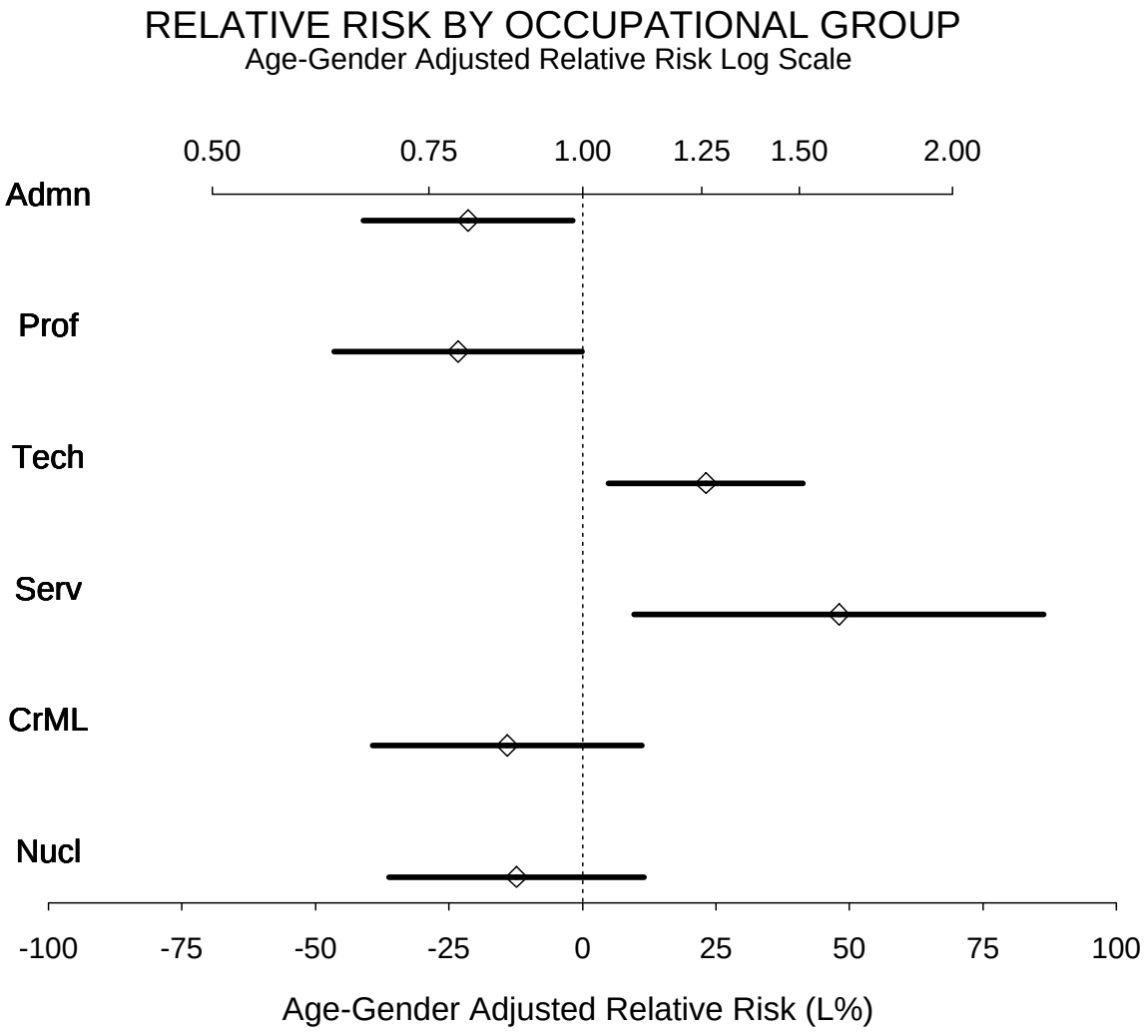

Figure 1: LRR estimates are shown in L% units (bottom horizontal scale). Point estimates are diamonds, and 95% CIs are horizontal lines for each occupational group.

References

[1] A. C. Atkinson. Likelihood ratios, posterior odds and information criteria. *J Econometrics*, 15:15–20, 1981.

[2] N. E. Breslow and N. E. Day. *Statistical Methods in Cancer Research, Volume II: The Design and Analysis of Cohort Studies*. Number Scientific Publication 82. International Agency for Research on Cancer, Lyon, 1987.

[3] J.M. Chambers and T.J. Hastie. *Statistical Models in S*. Wadsworth and Brooks/Cole, California, 1992.

[4] D. Clayton and M. Hills. *Statistical Models in Epidemiology*. Oxford University Press, 1993.

[5] N. Draper and H. Smith. *Applied Regression Analysis*. John Wiley and Sons, Inc., 2nd edition, 1981.

[6] B. Francis, M. Green, and C. Payne. *GLIM 4 The Statistical System For Generalized Linear Interactive Modelling*. Royal Statistical Society, Oxford University Press, 1993. Release 4.

[7] E. L. Frome. The analysis of rates using Poisson regression models. *Biometrics*, (39):665–674, 1983.

[8] E. L. Frome and H. Checkoway. Use of Poisson regression models in estimating incidence rates and ratios. *A. Journal of Epidemiology*, 121(2):309–323, 1985.

[9] E. L. Frome, D. L. Cragle, and R. W. McLain. Poisson regression analysis of the mortality amomg a cohort of World War II nuclear industry workers. *Radait. Res.*, (123):38–152, 1990.

[10] E. L. Frome, D. L. Cragle, J. P. Watkins, S. Wing, C. Shy, W. M. Tankersley, and C. M. West. A mortality study of employees of the nuclear industry in Oak Ridge, Tennessee. *Radait. Res.*, pages 64–80, 1997.

[11] E. L. Frome, M. K. Kutner, and J.J. Beauchamp. Regression analysis of Poisson distributed data. *J. Am. Statist. Assoc.*, 68:935–940, 1973.

[12] E. L. Frome and M. D. Morris. Evaluating goodness of fit of Poisson regression models in cohort studies. *The Am Statist*, 43(3):144–147, 1989.

[13] S Greenland and J.M. Sanders. Estimation of a common effect parameter from sparse follow-up data. *Biometrics*, (41):55–68, 1985.

[14] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2 edition, 1989.

[15] K.J. Rothman and J. D. Boice. *Epidemiologic Analysis with a Programmablea Calculator*. NIH,U. S. Government Printing Office, Washington, D.C., 1979. NIH 79-1649.

[16] Statistical Sciences, Inc., Seattle. *S-PLUS Reference Manual*, 1993. version 3.2.

[17] L. Tornquist, P. Vartia, and Y. Vartia. How should relative changes be measured ? *The American Statistician*, 39:43–46, 1985.