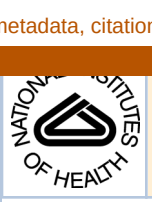

Published in final edited form as:

Radiat Environ Biophys. 2012 March ; 51(1): 15–22. doi:10.1007/s00411-011-0394-5.

Background stratified Poisson regression analysis of cohort data

David B. Richardson and

Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Bryan Langholz

Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Soto Street Building, Suite 202E 2001 N Soto Street, MC 9237, Los Angeles, CA 90089-9237, USA

David B. Richardson: david.richardson@unc.edu

Abstract

Background stratified Poisson regression is an approach that has been used in the analysis of data derived from a variety of epidemiologically important studies of radiation-exposed populations, including uranium miners, nuclear industry workers, and atomic bomb survivors. We describe a novel approach to fit Poisson regression models that adjust for a set of covariates through background stratification while directly estimating the radiation-disease association of primary interest. The approach makes use of an expression for the Poisson likelihood that treats the coefficients for stratum-specific indicator variables as 'nuisance' variables and avoids the need to explicitly estimate the coefficients for these stratum-specific parameters. Log-linear models, as well as other general relative rate models, are accommodated. This approach is illustrated using data from the Life Span Study of Japanese atomic bomb survivors and data from a study of underground uranium miners. The point estimate and confidence interval obtained from this 'conditional' regression approach are identical to the values obtained using unconditional Poisson regression with model terms for each background stratum. Moreover, it is shown that the proposed approach allows estimation of background stratified Poisson regression models of non-standard form, such as models that parameterize latency effects, as well as regression models in which the number of strata is large, thereby overcoming the limitations of previously available statistical software for fitting background stratified Poisson regression models.

Keywords

Cohort studies; Poisson regression; Ionizing radiation; Survival analysis

Introduction

Background stratified Poisson regression is an approach that has been used in the analysis of data derived from a variety of epidemiologically important studies of radiation-exposed populations, including uranium miners, nuclear industry workers, and atomic bomb

Correspondence to: David B. Richardson, david.richardson@unc.edu.

Electronic supplementary material The online version of this article (doi:10.1007/s00411-011-0394-5) contains supplementary material, which is available to authorized users.

survivors, as well as in studies of populations exposed to non-radiological hazards (Preston et al. 1987; Lubin et al. 1995, 2000; Cardis et al. 2005; Beane Freeman et al. 2009; Muirhead et al. 2009). In such settings, investigators have judged that background stratification offered a useful alternative to regression modeling as an approach to control for confounding by a set of measured covariates. Stratification by two or more covariates results in statistical adjustment for all product terms defined by the cross-classification of these factors. The approach allows an investigator to focus attention on parametric modeling of the exposure-disease association of primary interest, while obtaining covariate control via stratification. Moreover, the use of background stratification may encourage an investigator to avoid the use of a stepwise variable selection procedure that involves equivalence testing or significance testing (Greenland 1989; Maldonado and Greenland 1993), to reduce the set of covariates included in the final multivariable Poisson regression analysis and thereby reduce the influence of model selection on estimated standard errors.

A background stratified regression model may involve a large number of strata; reliable estimation of the coefficients and associated standard errors for the indicator variables distinguishing different strata may pose computational difficulties. In the above cited examples (Preston et al. 1987; Lubin et al. 1995, 2000; Cardis et al. 2005; Beane Freeman et al. 2009; Muirhead et al. 2009), a specialized program called AMFIT, which is part of the EPICURE software package, was used for fitting background stratified Poisson regression models (Preston et al. 1993). The AMFIT program implements background stratification of the regression model by the inclusion of multiplicative stratum-specific parameters and uses a Gauss–Seidel algorithm in order to avoid matrix inversion when estimating the coefficients for the indicator variables describing different strata. Standard errors are not computed for stratum-specific parameter estimates, and the values for these stratum-specific parameters are suppressed in the standard model output; however, the estimated standard errors for the regression model parameters that are not suppressed are adjusted to account for the estimation of the stratum-specific parameters.

In the present paper, we describe an alternative approach for fitting background stratified Poisson regression models that may be implemented using standard statistical software. We frame our approach in terms of fitting of general relative rate Poisson regression models that encompass log-linear model forms as well as alternatives, such as the linear excess relative rate model frequently used in radiation epidemiology. Rather than the standard Poisson likelihood, we propose maximizing an alternative expression for the likelihood that avoids estimation of the stratum-specific parameters by treating these as ‘nuisance’ terms, sometimes termed the ‘conditional’ Poisson likelihood (Cummings et al. 2003a, b). The ‘conditional’ Poisson likelihood has been used in the context of matched cohort data (Cummings et al. 2003a, b) and recurrent events, but to our knowledge, its use has not been described previously for the analysis of grouped survival data with control for confounding by background stratification. Unlike conditional maximum likelihood estimation in logistic and Cox regression analyses, which is necessary to avoid bias that can arise when there are small strata (Breslow and Day 1980), the ‘conditional’ Poisson analysis is not used here to avoid bias arising from small strata, but rather because it offers a useful approach for fitting background stratified Poisson regression models that avoid estimation of large numbers of stratum-specific parameters. This approach could be employed in highly stratified analyses so that the individuals contributing follow-up time and events in a stratum are tightly matched on covariates; however, if the strata are too fine, some may lack cases (or person-time), thus dropping the information from that stratum (Frome 1983; Pearl 2000).

The present approach allows investigators to overcome some limitations to the AMFIT program for fitting background stratified Poisson regression models. One limitation of the AMFIT program is that there are computational constraints to the number of strata that may

be included in a model. Our approach avoids estimation of the stratum-specific parameters and therefore has no such constraints; this may be an advantage in analyses in which regression models are stratified on large numbers of multilevel study covariates, as in a pooled analysis of cohort data (Lubin et al. 1995; Cardis et al. 2007). Another limitation of the AMFIT program is the inability to easily fit models that parameterize latency functions for protracted exposures (Langholz et al. 1999); the approach that is described here facilitates fitting of a wide range of Poisson regression models. Conveniently, this approach may be implemented using standard statistical software that provides the flexibility to specify general models, compute likelihood contributions, and maximize the likelihood function. SAS code is provided to illustrate the fitting of these models.

Materials and methods

Consider a cohort study in which incident events or cases have been ascertained over a period of follow-up. An analytical data structure can be generated for the purposes of Poisson regression analyses consisting of counts of person-time and events cross-classified by levels of explanatory variables (Allison 1995; Singer and Willett 2003). Necessarily, such a grouped data structure requires the categorization of any covariate that was originally measured on a continuous scale. Table 1 provides an illustrative tabulation of person-time and events cross-classified by levels of a binary explanatory variable of primary interest, *Z*, and two binary covariates, *A* and *B*. For purposes of notation, let *s* index the strata defined by the cross-classification of the covariates, *A* and *B*.

Exponential rate models are often used for Poisson regression analyses of grouped survival data; an exponential rate model with binary indicator variables for the four levels of *s*, S_1, …,S_4, and an explanatory variable *Z* can be expressed as $\lambda(\alpha_s, \beta) = \exp(\alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \alpha_4 S_4 + \beta_1 Z)$, where α_s are the strata and β the exposure model parameters. Poisson regression models typically are fitted by maximizing the unconditional likelihood. Let c_{sz} and P_{sz} denote the numbers of cases and the person-years observed at each level of the explanatory variable, *Z*, in stratum *s* of the covariates. The contribution of the *sz*th cell to the log likelihood is

$$l_{sz} = c_{sz} \ln(P_{sz} \lambda(\alpha_s, \beta)) - P_{sz} \lambda(\alpha_s, \beta). \quad (1)$$

The likelihood from the piecewise exponential model with categorical covariates for event-time data is formally the same as the likelihood treating the number of events in each cell as independent Poisson random variables with mean values given by the exponential rate times the person-time (Frome 1983; Breslow and Day 1987). Fitting an exponential rate Poisson regression model to the data in Table 1 yields an adjusted estimate of the change in log rate for a 1-unit increase in *Z*, as well as adjusted estimates of the log rate of outcome for each level of *A* and *B* (Table 3).

Other model forms have been proposed, notably general relative rate models of the form

$$\lambda(\alpha, \beta) = \exp(\alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \alpha_4 S_4) \phi(Z, \beta), \quad (2)$$

which encompasses the exponential rate model, $\phi(Z, \beta) = \exp(\beta Z)$, as well as linear excess relative rate models widely used in radiation epidemiology, $\phi(Z, \beta) = 1 + \beta Z$, and mixture models that combine features of linear and exponential functions (Thomas 1981).

Background stratification in Poisson regression

Consider again the grouped data structure of person-time and events shown in Table 1. Background stratification offers an approach to adjustment for the effects of a set of covariates that define strata of the regression model. As in (2), we consider a model with multiplicative stratum-specific parameters, α_s, and a relative rate function, $\phi(Z, \beta)$. We propose implementing a background stratified Poisson regression analysis by using another Poisson likelihood yielding the same inferences regarding β, as we shall show. Under this model, the contribution of the sth stratum of the data to the log likelihood is given by

$$l_s(\beta)=\ln\left(\prod_{z\in R_s}\phi(z;\beta)^{c_{sz}}\right)-c_s\left(\ln\left(\sum_{z\in R_s}P_{sz}\phi(z;\beta)\right)\right), \quad (3)$$

where R_s is the set of unique exposure (z) values in stratum s and c_s is the total number of cases in the sth stratum. Note that the coefficients for the stratum-specific effects, α_s, are not part of the expression for the likelihood.

Given the equivalence of this expression for the Poisson likelihood and Breslow's approximation for the partial likelihood for tied events in Cox regression, log-linear Poisson regression models are sometimes fitted using statistical procedures for Cox regression, for example in the analysis of matched cohort data (Cummings et al. 2003a, b). Further, noting that $c_s = \Sigma_z\, c_{sz}$, the expression for the Poisson likelihood shown in (3) is formally equivalent to a multinomial likelihood with probabilities $P_{sz}\phi(z; \beta)/\Sigma_z\, P_{sz}\phi(z; \beta)$ allowing for other software options for fitting.

To facilitate estimation, the typical analytical data structure for a Poisson regression analysis can be transformed so that it includes one observation per stratum of the analysis (i.e., one record representing all cases and person-years of observation in that stratum). Table 2 illustrates the transformation of the data structure shown in Table 1 to a data structure that includes one observation per stratum of the analysis. Since the number of cases, and unique covariate values, may vary from stratum to stratum, it is useful to summarize these values within each stratum. A SAS program to implement this transformation and summarization of a grouped data structure is provided as an electronic appendix.

Using the data structure in Table 2, a log-linear Poisson regression model with background stratification on *A* and *B* may be fitted, obtaining a single estimated parameter and associated 95% likelihood-based confidence interval (Table 3). This parameter corresponds to the estimated log of the rate ratio for a 1-unit change in *Z*, adjusted for *A* and *B*. The point estimate and confidence interval are identical to those values obtained via a standard Poisson regression analysis that included the main effect for *Z*, and indicator terms for each stratum defined by combinations of *A* and *B*. 'Appendix 1' provides sample code to fit these Poisson regression models using the SAS statistical package (SAS Institute Inc., Cary, NC).

When using the AMFIT statistical program, the estimation of background stratified models requires additional workspace to compute intermediate values. To obtain this workspace, the program will adjust the internal workspace by reducing the number of new variables that can be created. If there is not enough unused memory to handle the number of strata requested, an error is generated and the strata command is ignored. Our proposed approach has no such limitations on the number of strata, since these nuisance terms are conditioned out.

Maximizing the likelihood expressions shown in (1) and (3) consistently yields identical point estimates and confidence intervals for the exposure variable (given complete stratification by the confounding variables in each analysis). This can be shown quite

succinctly. With the standard expression for the Poisson likelihood given in (1) and a regression model as in (2), the log-likelihood contribution from stratum s is $l_s(\alpha_s, \beta) = c_s\alpha_s + \Sigma_{z\in R_s} c_{sz} \ln(P_{sz}\phi(z, \beta)) - \exp(\alpha_s)\Sigma_{z\in R_s} P_{sz}\phi(z, \beta)$.

The associated score functions are $U(\alpha_s) = c_s - \exp(\alpha_s) \Sigma_{z\in R_s} P_{sz}\phi(z; \beta)$ and $U(\beta)=\sum_s \sum_{z\in R_s} \left[c_{sz} \frac{\phi'(z;\beta)}{\phi(z;\beta)} \right] - \exp(\alpha_s) \sum_{z\in R_s} P_{sz}\phi'(z;\beta)$.

Setting $U(\alpha_s) = 0$, as obtained at the maximum likelihood estimates for α_s, implies that $\exp(\hat{\alpha_s}) = c_s/\Sigma_{z\in R_s} P_{sz}\phi(z; \beta)$. This expression may be plugged into the score functions for β yielding $U(\beta)=\sum_s \sum_{z\in R_s} \left[c_{sz} \frac{\phi'(z;\beta)}{\phi(z;\beta)} - c_s \frac{\Sigma_{z\in R_s} P_{sz}\phi'(z;\beta)}{\Sigma_{z\in R_s} P_{sz}\phi(z;\beta)} \right]$, which are precisely the score equations from the Poisson likelihood expression given in (3). Likelihood-based confidence intervals from the likelihood expression in (1) (plugging in the $\hat{\alpha}$ into the log-likelihood) differ from the log of the likelihood expression in (3) only by a constant, so that the likelihood-based confidence intervals will be identical. In 'Appendix 2', it is shown that estimated standard errors and Wald-type confidence intervals obtained from Poisson regression analyses maximizing either (1) or (3) are also identical (given regression analysis with complete stratification by the confounding variables).

The proposed approach is general enough to accommodate non-standard models, such as those in which an explanatory variable in the regression model is itself a function of one or more unknown parameters. One important example arises when modeling latency effects for a protracted exposure (Langholz et al. 1999). The present approach permits fitting of these types of background stratified Poisson regression models.

Results

Example one—Life Span Study of Japanese atomic bomb survivors

To illustrate the comparability of the proposed approach for estimation using the AMFIT program for background stratified Poisson regression analysis, the data from a recent analysis of the association between radiation dose and thyroid cancer incidence among female Japanese atomic bomb survivors who were aged 20 years or older at the time of the bombings in August, 1945, are used (Richardson 2009). The study included 241 thyroid cancers ascertained during the period 1958–1998 among women in the LSS. The primary exposure of interest was defined as weighted DS02 thyroid radiation dose, expressed in weighted Gray (Gy). We fitted Poisson regression models of the form $\lambda(c, a, e, l, h, d) = e^{\alpha_i}(1 + \delta d)$ where c, a, e, l, h, and d denote city, attained age, age-at-exposure, distal location, Adult Health Study membership, and dose, respectively, and the baseline rate of thyroid cancer was described by the stratum-specific parameters α_i that index the strata defined by the cross-classification of the covariates c, a, e, l, and h. Point estimates and 90% likelihood-based confidence intervals obtained via the SAS statistical package were compared with those obtained using the AMFIT module of the EPICURE software package.

The estimated association between radiation dose and thyroid cancer, with background stratification on the cross-classification of strata defined by city, attained age, age at exposure, distal location, and Adult Health Study membership was (β = 0.70, 90% CI: 0.20, 1.46). The point estimate and 90% confidence interval obtained from the background stratified Poisson regression model fitted using the SAS statistical package were identical (to the second decimal place) to the values obtained using the AMFIT program in the Epicure software package with background stratification on these covariates.

Example two—uranium miners data

To illustrate an extension of the proposed approach for background stratified Poisson regression analysis to an analysis in which a time-varying weight function is used to describe the modification of a cumulative dose–response association by time-since-exposure, we use data from a study of underground uranium miners (Hornung and Meinhardt 1987; Langholz et al. 1999). The Colorado Plateau cohort that we examined includes 2,704 white men employed in underground uranium mining operations between 1 January 1950 and 31 December 1960. Vital status was ascertained through 31 December 1990. The outcome of interest, lung cancer mortality, was defined by underlying cause of death; the study cohort includes 263 lung cancer deaths. The primary exposure of interest was defined as cumulative radon exposure, expressed in working-level months, and was computed for each worker as the product of the duration of employment in each job in a year by the estimated radon exposure rate for that job. Person-time and lung cancer deaths were cross-classified by categories of attained age (<50, 50 to <55, 55 to <60, 60 to <65, 65 to <70, 70 to <75, 75 to <80, or ≥80 years), calendar period (defined in 5-year categories from 1950–1955 to 1985–1990), and cumulative exposure. As in Langholz et al. (1999), cumulative exposure was partitioned into six exposure time-windows defined by the intervals 0 to <5, 5 to <10, 10 to <15, 15 to <20, 20 to <30, and 30+ years since exposure; within each window, exposure was categorized into five groups (<48, 48 to <154, 154 to <392, 392+ WLM). First, we estimated the association between cumulative radon dose and lung cancer mortality assuming a time-constant model, fitting a linear excess relative rate Poisson regression model of the form $\phi = (1 + \theta z)$ with background stratification on all covariates where z represents total cumulative exposure, $z=\sum_{j=1}^{6} x_j$, and x_1–x_6 represent the time-window-specific exposures. Next, we fit a model that included six terms for the six time-window-specific exposures, $\phi = (1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6)$. Last, we fitted a regression model that incorporated a bilinear latency function (i.e., a triangular weighting function consisting of two attached lines) (Langholz et al. 1999), corresponding to a regression model of the form: $\phi=\left(1+\delta\sum_{1}^{6} x_j \times \left[\left(\frac{t_j}{\alpha_1}\right) \mathrm{I}[0 \le t_j<\alpha_1]+\left(\frac{\alpha_2-t_j}{\alpha_2-\alpha_1}\right) \mathrm{I}[\alpha_1 \le t_j<\alpha_2]\right]\right)$, where t_j is the midpoint of each time-window, and I[‘logical function’] equals 1 if ‘logical function’ is true else 0.

Fitting a background stratified linear excess relative rate model for lifetime cumulative exposure (under a time-constant model) led to an estimated ERR/100 working-level months of 0.32 (95% CI: 0.18, 0.60). Next, we estimated model coefficients for six exposure time-windows; the estimated coefficients (0.11, 0.22, 0.44, 0.58, 0.34, and 0.15, respectively) were positive for each time-window but were imprecise. Last, we fitted a background stratified Poisson regression model that incorporated a bilinear latency function to describe effect modification by time-since-exposure. The fitted bilinear function obtains a maximal value 6.5 (SE = 3.72) years after exposure and then declines linearly to a null value 44.4 (SE = 2.66) years after exposure; the estimated ERR/100 WLM at its maximal value was 0.62 (95% CI: 0.23, 1.51).

Discussion

In the present paper, it is shown that by using an expression for the Poisson likelihood that treats the coefficients for stratum-specific indicator variables as ‘nuisance’ variables, an investigator conditions out the effects of the set of covariates that define the stratifying factors and obtains an adjusted estimate for the exposure effect of primary interest. The point estimate and confidence interval obtained from this ‘conditional’ regression approach are identical to the values obtained using unconditional Poisson regression with models terms for each background stratum. Therefore, this approach yields unbiased estimates of

association with proper confidence interval coverage. The proposed approach may be useful in settings in which the number of strata is large, as might occur in pooled analyses of cohort data (in which there are a large number of study sites). Under the approach, we propose there are no limitations on the number of strata as we employ an expression for the Poisson likelihood, sometimes termed the 'conditional Poisson' likelihood (Cummings et al. 2003a, b), that avoids the need to estimate coefficients for the stratum-specific indicator variables. Moreover, our proposed approach is general enough to facilitate fitting non-standard models, such as flexible models for modification of exposure effects by time-since-exposure.

Of course, background stratification is only one approach to covariate control in regression analyses. Another approach is to develop a parsimonious regression model to estimate the exposure-outcome association of primary interest while adjusting for a set of covariates (Frome 1983; Frome and Checkoway 1985). A stepwise variable selection approach to regression model development will often result in similar relative rate estimates to the values obtained via a background stratified Poisson regression analysis. The use of a stratified Poisson regression model may encourage an investigator to avoid use of a stepwise variable selection procedure. Since a variable and all possible product terms are entered, or removed, from a model, simultaneously, the background stratified Poisson regression approach is not conducive to standard stepwise variable selection procedures. Rather, it tends to encourage adjustment for a set of potential confounders selected based on considerations such as background knowledge regarding causal relationships between study variables. This may be fortuitous, as there is no evidence that use of stepwise variable selection procedures necessarily leads to better results than those obtained by adjusting for all well-measured confounders identified a priori (Greenland 2008; Weng et al. 2009).

We have previously shown how general relative rate regression models could be fitted using standard statistical software (Richardson 2008; Langholz and Richardson 2010). The present paper extends this for fitting models that accommodate background stratification of general relative rate Poisson regression models using standard statistical software. This should further facilitate the ability of investigators to replicate prior analyses that used specialized software to fit such models, as well as permit investigators to fit alternative models. Poisson regression methods offer a useful approach to adjustment for a set of model covariates in cohort analyses. This paper should facilitate wider use of Poisson regression methods and their application for background stratified analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project was supported by grant R01-CA117841 from the National Cancer Institute, National Institutes of Health.

References

Allison, PD. Survival analysis using SAS: a practical guide. SAS Institute; Cary: 1995.

Beane Freeman LE, Blair A, Lubin JH, Stewart PA, Hayes RB, Hoover RN, Hauptmann M. Mortality from lymphohematopoietic malignancies among workers in formaldehyde industries: the National Cancer Institute Cohort. J Natl Cancer Inst. 2009; 101(10):751–761. [PubMed: 19436030]

Breslow, N.; Day, NE. Statistical methods in cancer research: the analysis of case-control studies. IARC Scientific Publications; Lyon: 1980.

Breslow, NE.; Day, NE. Statistical methods in cancer research: the design and analysis of cohort studies. International Agency for Research on Cancer; Lyon: 1987.

Cardis E, Vrijheid M, Blettner M, Gilbert E, Hakama M, Hill C, Howe G, Kaldor J, Muirhead CR, Schubauer-Berigan M, Yoshimura T, Bermann F, Cowper G, Fix J, Hacker C, Heinmiller B, Marshall M, Thierry-Chef I, Utterback D, Ahn YO, Amoros E, Ashmore P, Auvinen A, Bae JM, Solano JB, Biau A, Combalot E, Deboodt P, Diez Sacristan A, Eklof M, Engels H, Engholm G, Gulis G, Habib R, Holan K, Hyvonen H, Kerekes A, Kurtinaitis J, Malker H, Martuzzi M, Mastauskas A, Monnet A, Moser M, Pearce MS, Richardson DB, Rodriguez-Artalejo F, Rogel A, Tardy H, Telle-Lamberton M, Turai I, Usel M, Veress K. Risk of cancer after low doses of ionising radiation: retrospective cohort study in 15 countries. Br Med J. 2005; 331(7508):77. [PubMed: 15987704]

Cardis E, Vrijheid M, Blettner M, Gilbert E, Hakama M, Hill C, Howe G, Kaldor J, Muirhead CR, Schubauer-Berigan M, Yoshimura T, Bermann F, Cowper G, Fix J, Hacker C, Heinmiller B, Marshall M, Thierry-Chef I, Utterback D, Ahn YO, Amoros E, Ashmore P, Auvinen A, Bae JM, Solano JB, Biau A, Combalot E, Deboodt P, Diez Sacristan A, Eklof M, Engels H, Engholm G, Gulis G, Habib R, Holan K, Hyvonen H, Kerekes A, Kurtinaitis J, Malker H, Martuzzi M, Mastauskas A, Monnet A, Moser M, Pearce MS, Richardson DB, Rodriguez-Artalejo F, Rogel A, Tardy H, Telle-Lamberton M, Turai I, Usel M, Veress K. The 15-country collaborative study of cancer risk among radiation workers in the nuclear industry: estimates of radiation related cancer risks. Radiat Res. 2007; 167(4):396–416. [PubMed: 17388693]

Cummings P, McKnight B, Greenland S. Matched cohort methods for injury research. Epidemiol Rev. 2003a; 25:43–50. [PubMed: 12923989]

Cummings P, McKnight B, Weiss NS. Matched-pair cohort methods in traffic crash research. Accid Anal Prev. 2003b; 35(1):131–141. [PubMed: 12479904]

Frome EL. The analysis of rates using Poisson regression models. Biometrics. 1983; 39(3):665–674. [PubMed: 6652201]

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(2):309–323. [PubMed: 3839345]

Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health. 1989; 79(3):340–349. [PubMed: 2916724]

Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. Am J Epidemiol. 2008; 167(5):523–529. discussion 530–521. [PubMed: 18227100]

Hornung RW, Meinhardt TJ. Quantitative risk assessment of lung cancer in U.S. uranium miners. Health Phys. 1987; 52(4):417–430. [PubMed: 3032855]

Langholz B, Richardson DB. Fitting general relative risk models for survival time and matched case-control analysis. Am J Epidemiol. 2010; 171(3):377–383. [PubMed: 20044379]

Langholz B, Thomas D, Xiang A, Stram D. Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado Plateau uranium miners cohort. Am J Ind Med. 1999; 35(3):246–256. [PubMed: 9987557]

Lubin JH, Boice JD Jr, Edling C, Hornung RW, Howe GR, Kunz E, Kusiak RA, Morrison HI, Radford EP, Samet JM, et al. Lung cancer in radon-exposed miners and estimation of risk from indoor exposure. J Natl Cancer Inst. 1995; 87(11):817–827. [PubMed: 7791231]

Lubin JH, Pottern LM, Stone BJ, Fraumeni JF Jr. Respiratory cancer in a cohort of copper smelter workers: results from more than 50 years of follow-up. Am J Epidemiol. 2000; 151(6):554–565. [PubMed: 10733037]

Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol. 1993; 138(11):923–936. [PubMed: 8256780]

Muirhead CR, O'Hagan JA, Haylock RG, Phillipson MA, Willcock T, Berridge GL, Zhang W. Mortality and cancer incidence following occupational radiation exposure: third analysis of the National Registry for Radiation Workers. Br J Cancer. 2009; 100(1):206–212. [PubMed: 19127272]

Pearl, J. Causality: models, reasoning, and inference. Cambridge University Press; Cambridge: 2000.

Preston DL, Kato H, Kopecky KJ, Fujita S. Studies of the mortality of A-bomb survivors, report 8. Cancer mortality, 1950–1982. Radiat Res. 1987; 111(1):151–178. [PubMed: 3446217]

Preston, DL.; Lubin, JH.; Pierce, DA.; McConney, ME. Epicure: user's guide. Hirosoft International Corporation; Seattle: 1993.

Richardson DB. A simple approach for fitting linear relative rate models in SAS. Am J Epidemiol. 2008; 168(11):1333–1338. [PubMed: 18953061]

Richardson DB. Exposure to ionizing radiation in adulthood and thyroid cancer incidence. Epidemiology. 2009; 20(2):181–187. [PubMed: 19177023]

Singer, JD.; Willett, JB. Applied longitudinal data analysis: modeling change and event occurrence. Oxford University Press; New York: 2003.

Thomas D. General relative risk models for survival time and matched case-control analysis. Biometrics. 1981; 37(4):673–686.

Weng HY, Hsueh YH, Messam LL, Hertz-Picciotto I. Methods of covariate selection: directed acyclic graphs and the change-inestimate procedure. Am J Epidemiol. 2009; 169(10):1182–1190. [PubMed: 19363102]

Appendix 1

A standard log-linear unconditional Poisson regression model of the form $\lambda(\alpha, \beta) = \exp(\alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \alpha_4 S_4 + \beta_1 Z)$ may be fitted to the data in Table 1 via the SAS statistical package as follows:

```
proc nlp data= ;
parms a1-a4 b1 ;
profile a1-a4 b1/ alpha=0.05 forchi=chi;
lambda=exp(a1*S1 + a2*S2 + a3*S3 + a4*S4 + b1*Z) ;
LL=c*log(P*lambda) - (P*lambda) ;
max LL ; run;
```

The variables *P* and *c* denote counts of person-time and events, respectively, in the grouped data structure. The 'parms' statement defines the parameters to be estimated, and the 'profile' statement requests associated 95% likelihood-based confidence intervals. The term 'lambda' specifies that the rate of disease conforms to an exponential function of the model covariates. The 'LL' statement specifies the expression for the unconditional Poisson likelihood, and the statement 'max LL' defines the function to be maximized.

A log-linear Poisson regression model may be fitted to the data structure in Table 2, with background stratification on covariates *A* and *B*, via the SAS statistical package as follows:

```
proc nlp data= ;
parms b1 ;
profile b1 / alpha=0.05 forchi=chi;
array _cases{*} _cases1-_cases2;
array _pt{*} _pt1-_pt2;
array _z{*} _z1-_z2;
caseprod=1; sum=0; nc=_ncovals;
do i = 1 to nc;
phi = exp(b1*_z{i}) ;
caseprod = caseprod*phi**_cases{i} ;
sum = sum+phi*_pt{i} ;
end;
```

```
LL= log(caseprod) - _totcases * log(sum) ;
max LL; run;
```

The analytical data structure has one record per stratum. The variables *_ncovals* and *_totcases* denote the total number of exposure values, and total number of cases, in each stratum. The arrays *_cases*, *_pt*, and *_z* index the values for the counts of events, person-time, and levels of the exposure variable(s) of interest in each stratum of the analytical data structure. The length of the arrays will depend upon the analytical data structure. The variables *caseprod* and *sum*, which are the numerator and denominator, respectively, of the expression for the conditional likelihood, are initialized at each new record in the analysis. The term 'phi' defines the relative rate function of the regression model. In the example above, the rate ratio function conforms to a standard log-linear model. The 'parms' statement defines the parameter(s) to be estimated, and the 'profile' statement requests associated 95% profile likelihood confidence bounds. The 'LL' statement specifies the expression for the log likelihood in this model, and the statement 'max LL' defines the function to be maximized.

The SAS procedure PROC NLP is part of the SAS/OR statistical package. Some SAS users may have access to the SAS/STAT package but not the SAS/OR package. Therefore, below, we also provide sample code for fitting background stratified Poisson regression models via the SAS PROC NLMIXED procedure which is part of the SAS/STAT package. SAS PROC NLMIXED does not directly output profile likelihood confidence intervals for estimated parameters but does report Wald-type confidence intervals.

```
proc nlmixed data= ;
parms b1=0;
array _cases{*} _cases1-_cases2;
array _pt{*} _pt1-_pt2;
array _z{*} _z1-_z2;
caseprod = 1; sum=0;
do i = 1 to _ncovals;
phi= exp(b1*_z{i});
caseprod = caseprod*phi**_cases{i};
sum = sum+phi*_pt{i};
end;
LL= log(caseprod) - _totcases * log(sum) ;
model _totcases ~ general(LL);
run;
```

This approach accommodates a variety of functional forms for the relative rate function, φ. For example, a linear excess relative rate model of the form $\varphi = (1 + \beta z)$ would be fitted by replacing the statement "phi= exp (b1*_z{i});" with the statement "phi= (1+b1*_z{i});".

Appendix 2

With the model as in (2), the unconditional log-likelihood contribution from stratum s is $L_s(\alpha_s, \beta) = c_s\alpha_s + \Sigma_{z\in R_S} c_{sz} \ln(P_{sz}\phi(z, \beta)) - \exp(\alpha_s) \Sigma_{z\in R_S} P_{sz}\phi(z, \beta)$. The observed information at $\hat{\alpha}$, $I(\hat{\alpha}, \hat{\beta})$ is given by

$$-dU(\widehat{\alpha}_s)/d\alpha_s = c_s$$
$$-dU(\widehat{\alpha}_s)/d\alpha_l = 0$$
$$-dU(\widehat{\alpha}_s)/d\beta = c_s \sum_{z\in R_s} P_{sz}\phi'(z;\beta)/\sum_{z\in R_s} P_{sz}\phi(z;\beta)$$
$$-dU(\beta)/d\beta = \sum_s \left[\sum_{i\in D_s} \left[\frac{\phi''(Z_i;\beta)}{\phi(Z_i;\beta)} - \left(\frac{\phi'(Z_i;\beta)}{\phi(Z_i;\beta)} \right)^2 \right] - c_s \frac{\sum_{z\in R_s} P_{sz}\phi''(z;\beta)}{\sum_{z\in R_s} P_{sz}\phi(z;\beta)} \right].$$

The variance estimate for $\hat{\beta}$ is the corner of the inverse of the observed information (evaluated at $\hat{\alpha}, \hat{\beta}$) which may be obtained using the well-known matrix formula

$$(I^{-1})_{\beta,\beta} = \left[I_{\beta,\beta} - I_{\beta,\alpha} I_{\alpha,\alpha}^{-1} I_{\alpha,\beta} \right]^{-1}.$$

Since $I_{\alpha,\alpha}$ is a diagonal matrix, it is easy to compute that

$$\mathrm{var}(\widehat{\beta})^{-1} = \sum_s \left[\sum_{i\in D_s} \left[\frac{\phi''(Z_i;\beta)}{\phi(Z_i;\beta)} - \left(\frac{\phi'(Z_i;\beta)}{\phi(Z_i;\beta)} \right)^2 \right] - c_s \left[\frac{\sum_{z\in R_s} P_{sz}\phi''(z;\beta)}{\sum_{z\in R_s} P_{sz}\phi(z;\beta)} - \left(\frac{\sum_{z\in R_s} P_{sz}\phi'(z;\beta)}{\sum_{z\in R_s} P_{sz}\phi(z;\beta)} \right)^2 \right] \right]$$

This expression is the same as the second derivative of the ‘conditional’ Poisson log-likelihood; consequently, estimated standard errors and associated Wald-type confidence intervals will be the same. For simplicity, we derive the expression for a single parameter β. The expressions apply to column vector $\boldsymbol{\beta}$ where the derivatives are as in standard vector calculus and squared terms are replaced by outer products, i.e., replace a^2 by $\boldsymbol{a}\boldsymbol{a}^t$ where $\boldsymbol{a}^t$ is the transpose of $\boldsymbol{a}$.

Table 1

Hypothetical cohort data

Observation	*A*	*B*	*Z*	Cases	Person-years	Stratum, *s*
1	0	0	0	21	1,325	1
2	0	0	1	32	2,362	1
3	0	1	0	13	353	2
4	0	1	1	21	1,322	2
5	1	0	0	10	226	3
6	1	0	1	2	1,141	3
7	1	1	0	11	111	4
8	1	1	1	4	1,042	4

Data structure representing the numbers of cases and person-years at risk, cross-classified by categories of exposure, *Z*, and covariates, *A* and *B*

Table 2

Transformation of the data shown in Table 1 into a data structure, representing the numbers of cases and person-years at risk at each unique value of the explanatory variable, *Z*, within strata defined by covariates *A* and *B*

Stratum, *s*	c_{s0}	P_{s0}	c_{s1}	P_{s1}
1	21	1,325	32	2,362
2	13	353	21	1,322
3	10	226	2	1,141
4	11	111	4	1,042

Table 3

Estimated parameters, associated standard errors, and 95% likelihood-based confidence intervals (CI) obtained by fitting an unconditional Poisson regression model to the data in Table 1 and by fitting a conditional Poisson regression model (with background stratification on covariates *A* and *B*) to the data shown in Table 2

Parameter	Unconditional		Conditional	
	Estimate	95% CI	Estimate	95% CI
α_1	−3.706	−4.027, −3.412	–	–
α_2	−3.181	−3.601, −2.800	–	–
α_3	−3.958	−4.627, −3.385	–	–
α_4	−3.462	−4.085, −2.910	–	–
β_1	−1.043	−1.425, −0.659	−1.043	−1.425, −0.659