AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 121, No. 2
Printed in U.S.A.
Copyright © 1985 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

# Epidemiologic Programs for Computers and Calculators

## USE OF POISSON REGRESSION MODELS IN ESTIMATING INCIDENCE RATES AND RATIOS

EDWARD L. FROME[1] AND HARVEY CHECKOWAY[2]

**Frome, E. L. (Oak Ridge National Laboratory, Oak Ridge, TN 37830) and H. Checkoway. Use of Poisson regression models in estimating incidence rates and ratios. *Am J Epidemiol* 1985;121:309–23.**

**Summarizing relative risk estimates across strata of a covariate is commonly done in comparative epidemiologic studies of incidence or mortality. Conventional Mantel-Haenszel and rate standardization techniques used for this purpose are strictly suitable only when there is no interaction between relative risk and the covariate, and tests for interaction typically are limited to examination for departures from linearity. Poisson regression modeling offers an alternative technique which can be used for summarizing relative risk and for evaluating complex interactions with covariates. A more general application of Poisson regression is its utility in modeling disease rates according to postulated etiologic mechanisms of exposures or according to disease expression characteristics in the population. The applications of Poisson regression analysis to problems of summarizing relative risk and disease rate modeling are illustrated with examples of cancer incidence and mortality data, including an example of a nonlinear model predicted by the multistage theory of carcinogenesis.**

**biometry; epidemiologic methods; Poisson distribution; regression analysis**

Incidence or mortality data obtained from epidemiologic follow-up studies are often expressed as covariate stratum-specific rates, for which the covariate may be age or some other presumed confounding factor. Comparative analysis of rates between groups characterized according to exposure level commonly involves computation of summary relative risk (or rate ratio) estimates, wherein stratum-specific relative risks are condensed into a single measure (1). Several methods have been proposed for this purpose. The Mantel-Haenszel summary odds ratio method (2) has been adapted for use in cohort studies by Rothman and Boice (3) and by Tarone (4). Lilienfeld and Pyne (5) have also proposed summarizing relative risk with a procedure which differs from the previous approaches principally in the choice of weights assigned to the stratum-specific relative risk estimates. Miettinen (6) has discussed this problem in the context of standardization as a means of controlling confounding.

Received for publication May 19, 1983, and in final form May 15, 1984.

[1] Mathematics and Statistics Research, Engineering Physics and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN.

[2] Department of Epidemiology, University of North Carolina, School of Public Health, Chapel Hill, NC.

Reprint requests to Dr. Edward L. Frome, Oak Ridge National Laboratory, P.O. Box Y, Oak Ridge, TN 37830.

This report is based on work performed under Contract DE-AC05-76OR00033 with the US Department of Energy, Office of Health and Environmental Research, and the Oak Ridge Associated Universities, and Contract W-7405-eng-26 with the Union Carbide Corporation.

The authors are grateful to Drs. Everett Logue, Elaine Zeighami, D. G. Gosslee, and M. D. Morris for offering helpful suggestions, and to Pamela Hooker for help with the manuscript.

With these relative risk summarizing techniques, there is the implicit assumption that there is no interaction between relative risk and levels of the covariates, and as such, the comparison of summary estimators is strictly valid only when this condition is true (7). Testing for interaction typically is limited to examining for departures from linearity of relative risk; however, when there are complex relationships between disease rates and the covariate, conventional summary estimation techniques are unsuitable. An alternative approach to this problem is offered by modeling disease rates as a function of covariate levels. Modeling disease rates can also be used for more general purposes, such as to describe the pattern of disease occurrence according to postulated etiologic mechanisms of exposure, e.g., initiation and promotion activities of carcinogens, or according to disease expression characteristics in the population, e.g., bimodal age distribution.

In this paper, we present a method for obtaining summary relative risk estimates by means of Poisson regression modeling. Examples are presented to illustrate the method of summarizing relative risk and to demonstrate how Poisson regression can be used to address problems of heterogeneity and interaction. An example that uses a nonlinear disease rate model derived from the multistage theory of carcinogenesis is given to illustrate the informativeness of the Poisson regression modeling approach.

### *The Poisson regression model*

The general framework of the Poisson regression model, as applied to situations in which the dependent variable is a count (e.g., number of incident cases in a cohort study), has been described previously by Frome et al. (8). The reader is referred to reference 9 for a discussion of the application of the model for epidemiologic analysis of cohort data. To specify a Poisson regression model, it is assumed that the dependent variable follows the Poisson distribution and that a rate function $\lambda(X, \beta)$ that describes the relationship between disease rates, the predictor variables ($X$), and the unknown vector of parameters ($\beta$) is given.

In the context of a cohort study, a population for which incidence or mortality data for some disease have been obtained can be categorized into $J$ strata of the covariate (e.g., age) for each of $K$ "risk" (exposure) groups. The data can be summarized for each level of these two factors as shown in table 1, where $y_{jk}$ denotes the number of cases and $c_{jk}$ the number of persons or person-years for risk group $k$ in stratum $j$. The corresponding observed disease rate is $r_{jk} = y_{jk}/c_{jk}$. If risk group 1 is considered to be the reference, or nonexposed, group, the estimated relative risk for group $k$ ($k = 2, \ldots, K$) for stratum $j$ ($j = 1, \ldots, J$) is $r_{jk}/r_{j1}$.

It is assumed that the $y_{jk}$ are distributed as Poisson variates (10, 11) with expectation $\mu_{jk} = c_{jk}\lambda_{jk}$. The $\lambda_{jk}$ represent the underlying rate functions which are estimated from the data by $r_{jk}$. If the covariate stratum-specific relative risks are constant within each risk group, then $\lambda_{jk} = \lambda_j\phi_k$, where $\lambda_j$ denotes the rate for the $j$th stratum level, $\phi_k$ is the summary relative risk for group $k$ ($k > 1$), and $\phi_k = 1$ for group $k = 1$. This is referred to as the *product model*, and for estimation purposes, it can be reparameterized as a log-linear model, i.e.,

$$\lambda_{jk} = \exp(\alpha_j + \delta_k), \qquad (1)$$

where $\alpha_j = \ln(\lambda_j)$ and $\delta_k = \ln\phi_k$ ($k = 2, \ldots,$

TABLE 1
*Data layout for analysis of cohort study data*

| Stratum | | Risk group | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | ... | $k$ | ... | $K$ |
| 1 | Cases | $y_{11}$ | $y_{12}$ | ... | $y_{1k}$ | ... | $y_{1K}$ |
| | Person-years | $c_{11}$ | $c_{12}$ | ... | $c_{1k}$ | ... | $c_{1K}$ |
| . | | . | . | | . | | . |
| . | | . | . | | . | | . |
| . | | . | . | | . | | . |
| $j$ | Cases | $y_{j1}$ | $y_{j2}$ | ... | $y_{jk}$ | ... | $y_{jK}$ |
| | Person-years | $c_{j1}$ | $c_{j2}$ | ... | $c_{jk}$ | ... | $c_{jK}$ |
| . | | . | . | | . | . | |
| . | | . | . | | . | . | |
| . | | . | . | | . | . | |
| $J$ | Cases | $y_{J1}$ | $y_{J2}$ | ... | $y_{Jk}$ | ... | $y_{JK}$ |
| | Person-years | $c_{J1}$ | $c_{J2}$ | ... | $c_{Jk}$ | ... | $c_{JK}$ |

$K$). The $\alpha_j$ correspond to the natural logarithms of the stratum-specific incidence rates in the reference group, while the $\delta_k$ represent the natural logarithm of the summary relative risk for group $k$ (with group 1 as the reference group).

Tarone (4) has discussed the special case that occurs when there are only two risk groups, and the more general case has been considered by Miettinen (6). Two of their examples will be given to illustrate how Poisson regression can be used to implement and extend their analyses.

*Example 1—Relative risk estimation for two risk groups*

This example illustrates how Poisson regression can be used when there are only two risk groups. In this situation, $\lambda_{j1} = \lambda_j$ and $\lambda_{j2} = \lambda_j\phi = \exp(\alpha_j + \delta)$. Scotto et al. (12) compared skin cancer incidence rates among women for two geographic areas, Dallas-Fort Worth and Minneapolis-St. Paul. Tarone (4) previously derived a summary relative risk for Dallas-Fort Worth from the age-stratified data shown in table 2. The formula Tarone used for the summary relative risk, in notation consistent with formulae given here, is

$$\hat{\phi} \simeq \left(\sum \frac{c_{j1}}{c_j} y_{j2}\right) \Big/ \left(\sum \frac{c_{j2}}{c_{j.}} y_{j1}\right), \qquad (2)$$

TABLE 2

*Number of cases of nonmelanoma skin cancer and population size for women in Dallas-Fort Worth and Minneapolis-St. Paul**

| Age group (years) | Minneapolis-St. Paul | | Dallas-Fort Worth | | Relative risk† |
|---|---|---|---|---|---|
| | No. of cases | Population size | No. of cases | Population size | |
| 15–24 | 1 | 172,675 | 4 | 181,343 | 3.81 |
| 25–34 | 16 | 123,065 | 38 | 146,207 | 2.00 |
| 35–44 | 30 | 96,216 | 119 | 121,374 | 3.14 |
| 45–54 | 71 | 92,051 | 221 | 111,353 | 2.57 |
| 55–64 | 102 | 72,159 | 259 | 83,004 | 2.21 |
| 65–74 | 130 | 54,722 | 310 | 55,932 | 2.33 |
| 75–84 | 133 | 32,185 | 226 | 29,007 | 1.89 |
| 85+ | 40 | 8,328 | 65 | 7,538 | 1.80 |

* Adapted from Scotto et al. (12).

† With Minneapolis-St. Paul as the reference group.

where $c_{j.} = \sum_k c_{jk}$ = total number of persons in the $j$th age stratum, summed across $K$ risk groups. Tarone computed a summary relative risk of 2.24 (standard deviation (SD) = 0.12) for these data, with a $\chi^2$ for heterogeneity of 8.22 with 7 df, indicating uniformity of relative risk across strata.

The Poisson regression model for these data (see Appendix for a discussion of the general case) can be expressed as

$$\lambda(X_i, \boldsymbol{\beta}) = \exp(X_i\boldsymbol{\beta})$$

$$= \exp\left(\sum_j x_{ij}\beta_j\right) = \exp(\alpha_j + \delta), \qquad (3)$$

where $X_i = (x_{i1}, \ldots, x_{i9})$ is a row vector and $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_8, \delta)'$. In this example, $x_{i1}$ through $x_{i8}$ are indicator variables representing terms for each age stratum, and $x_{i9}$ indicates risk group, so that $\delta = \ln(\phi)$, where $\phi$ denotes the summary relative risk for Dallas-Fort Worth with Minneapolis-St. Paul as the reference group. Note that $X_i$ is the $i$th row of the 16 by 9 model matrix $\boldsymbol{X}$. In general, for a $J$ by $K$ table, the model matrix will be $JK$ by $J + K - 1$. A detailed discussion of a procedure used in constructing a model matrix is given in the Appendix. The maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ of the parameter vector $\boldsymbol{\beta}$ is obtained using the iteratively reweighted least squares method desribed by Frome (9). The computations can be performed using the Generalized Linear Iterative Models (GLIM) statistical package (13). (Listings of the computer program used for these analyses can be obtained from the author (E. L. F.) on request.)

The maximum likelihood estimate of $\delta$ is $\hat{\delta} = 0.804$ (the unadjusted estimate is $\hat{\delta} = 0.743$), from which $\hat{\phi} = 2.23$ is computed. This result agrees closely with that obtained from equation 2 by Tarone (4). The estimated standard deviation for $\hat{\delta}$ is obtained from the inverse of the information matrix, which is readily available following the last iteration of the iteratively reweighted least squares procedure. The computed standard deviation is 0.0522, which

gives a 95 per cent confidence interval for $\hat{\phi}$ of 2.02–2.48. Heterogeneity of relative risk across strata is evaluated from the deviance $D(\hat{\beta})$, which is a measure of unexplained variation (see Appendix). In this example, $D(\hat{\beta}) = 8.17$ with 7 df, indicating that the product model cannot be rejected, i.e., it is reasonable to assume that the relative risk is constant across age strata.

A plot of the logarithm of the incidence rates against the log of age − 15 years (where age is the midpoint of the stratum-specific interval in years) reveals that there is additional structure in these data (see figure 1). Figure 1 suggests that $\ln\lambda$, the natural logarithm of the age-specific rate, increases linearly with $\ln t$, and the parameter $\theta$ is the slope of the line. This is the relationship between incidence and age ($t$) that is predicted by the multistage theory of carcinogenesis (14) and is referred to as the *power law* because the age-specific incidence rates are proportional to $t^{\theta}$. For the data in table 2—$\lambda_{j1} = \lambda_0 t^{\theta}$ and $\lambda_{j2} = \lambda_0 t^{\theta}\phi$, which can be written as

$$\lambda(X_i, \beta) = \exp(\alpha x_{i1} + \theta x_{i2} + \delta x_{i3}), \quad (4)$$

where $x_{i1} = 1$, $x_{i2} = \ln(t_i)$, and $x_{i3} = 1$ for Dallas-Fort Worth and 0 otherwise. If $t$ = (age − 15 years)/35, then $t = 1$ for the 45–54 age stratum, and the "intercept" term in equation 4 is the logarithm of the incidence rate for Minneapolis-St. Paul, i.e., $j = 4$ (ages 45–54), $x_{i2} = x_{i3} = 0$, and, from equation 4, $\lambda_0 = \exp(\alpha)$. For Dallas-Fort Worth, when $j = 4$ and $t = 1$, $\lambda_{j2} = \exp(\alpha + \delta) = \lambda_0\phi$. The maximum likelihood estimates for the parameters in equation 4 are $\hat{\alpha} = -0.168$ (SD = 0.048), $\hat{\theta} = 2.29$ (SD = 0.063), and $\hat{\delta} = 0.803$ (SD = 0.052). The value for deviance, which is asymptotically distributed as a $\chi^2$ (8), is 14.36 with 13 df, indicating a good fit of the power model for these data.

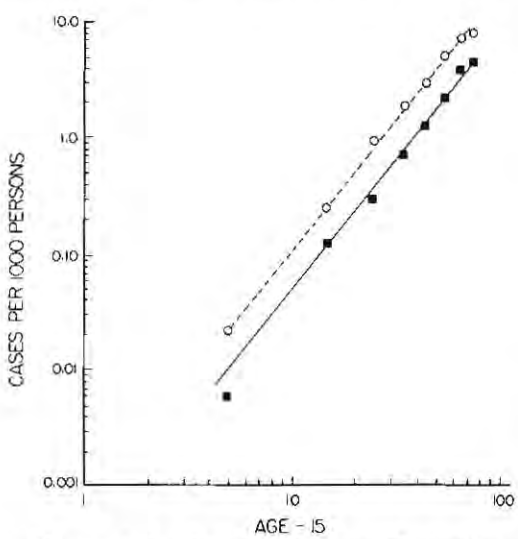


FIGURE 1. Log-log plot of age-specific incidence rates for skin cancer data for Dallas-Fort Worth (O- - -O) and Minneapolis-St. Paul (■- ■) (12).

The findings from the analysis described above are summarized in a Poisson ANOVA (analysis of variance) format in table 3. This table is obtained by recording the value of the deviance and the degrees of freedom for each model considered and can be used to evaluate the importance of parameters in the model and the goodness of fit of different models. The deviance provides an *absolute* measure of residual (unexplained) variation and can be compared with the chi-square distribution to assess the goodness of fit of a specific model. (The mean and variance of a chi-square statistic are the degrees of freedom and two times the degrees of freedom, respectively.)

As discussed previously, the deviance of

TABLE 3
*Poisson ANOVA table for skin cancer incidence data in table 2*

| Model | $\ln(\lambda)$ | No. of parameters | Deviance* | df |
|---|---|---|---|---|
| Minimal | $\alpha$ | 1 | 2,790.3 | 15 |
| Power law ($\delta = 0$) | $\alpha + \theta t$ | 2 | 272.7 | 14 |
| Age alone | $\alpha_j$ | 8 | 266.9 | 8 |
| City alone | $\alpha + \delta$ | 2 | 2,569.7 | 14 |
| Power law | $\alpha + \theta t + \delta$ | 3 | 14.4 | 13 |
| Age and city | $\alpha_j + \delta$ | 9 | 8.2 | 7 |
| Complete | $\alpha_{jk}$ | 16 | 0.0 | |

* The deviance provides an absolute measure of residual (i.e., unexplained) variation and is asymptotically distributed as a chi-square random variable (see Appendix).

8.2 (df = 7) for the product model indicates that the relative risk is constant across age strata. The deviance of the power law model (14.4, df = 13) further suggests a good fit of a model predicted by the multistage carcinogenesis theory.

Further inspection of table 3 indicates that there is considerable lack of fit for all other models considered in that the deviance is considerably larger than the corresponding degrees of freedom for each of these models. In this situation, the more parsimonious power law model is judged to provide a "better" representation of these data. A more formal way to make this evaluation is to consider the difference of the deviances 14.4 − 8.2 = 6.2. This is a likelihood ratio statistic for the model $\alpha_j = \alpha + \theta \ln t$ and can be compared with the chi-square distribution with 13 − 7 = 6 df. Comparison of the corresponding differences in the deviances and associated degrees of freedom suggests that the power model provides a good fit to the data.

The importance of the parameter $\delta$ can be evaluated by computing the decrease in the deviance that occurs when $\delta$ is included in the power law model (272.7 − 14.4) which is 258.3 with 1 df. This is consistent with the information obtained from the point estimate and its estimated standard deviation ($\hat{\delta}^2/\mathrm{SD}\hat{\delta}^2 = 238.5$), i.e., $\hat{\delta}$ exceeds the null value of zero by more than 15 standard deviations. The deviance can also be used to construct an $R^2$-type measure of variance explained. For example, the first three lines of table 3 show that most of the variation, 100(2,790.3 − 266.9)/2,790.3 = 90.4 per cent, in these data is explained by age alone, and that the power law model (with $\delta = 0$) accounts for most (90.2 per cent, i.e., 100(2,790.3 − 272.7)/2,790.3) of this explained variation. This example illustrates that the relative importance of various parameters can be assessed in several ways using the Poisson ANOVA table and that these results are consistent with results obtained when point and interval estimates of summary relative risks are used. We emphasize that this ad hoc analysis of the power law model should be interpreted as such, and that a formal statistical test of the goodness of fit of a model that is obtained after inspecting the data is of limited value. The difference of the deviance should be used to assess the *relative* importance of a specific set of parameter values and is the analog of the decrease in sum of squares for standard linear models. This implies that one should interpret the Poisson ANOVA in the same manner as an ANOVA table for an unbalanced design, and it should be recognized that the results are order-dependent. The general procedure that we recommend is to enter the potential confounding factor first and then to inspect the decrease in the deviance that occurs when the risk factor of interest is entered into the model, as well as the deviance for the model with both factors included.

In summary, the analysis for this example indicates that the relative risk for skin cancer among women is significantly higher in Dallas-Fort Worth than in Minneapolis-St. Paul, that the relative risk is constant with respect to age, and that the age-specific incidence rates are proportional to age − 15 years raised to the power of 2.3.

*Example 2—Relative risk estimation for K risk group*

This next example is presented to demonstrate how Poisson regression modeling can be used to estimate the relative risk when there are more than two risk groups. Miettinen (6) has discussed the general problem of summarizing relative risk in the context of standardization with respect to the distribution of a confounder. The method of risk ratio standardization that Miettinen (6) used was illustrated with data from a cohort study by Kahn (15) of lung cancer mortality in relationship to cigarette smoking. Table 4 shows the lung cancer death rates according to the levels of the risk factor and age, the potential confounder.

The crude relative risk (i.e., ignoring age

TABLE 4
*Lung cancer mortality according to cigarette consumption and age for current cigarette smokers**

| Age group (years) | | Nonsmokers | Cigarettes/day | | | | |
|---|---|---|---|---|---|---|---|
| | | | Occasional | 1–9 | 10–20 | 21–39 | 40+ |
| 35–44 | Deaths | 0 | 0 | 0 | 2 | 4 | 0 |
| | Person-years | 35,164 | 3,657 | 8,063 | 59,965 | 40,643 | 3,992 |
| | (Rate per $10^5$) | (0) | (0) | (0) | (3.34) | (9.84) | (0) |
| 45–54 | Deaths | 0 | 0 | 0 | 2 | 10 | 2 |
| | Person-years | 15,134 | 1,283 | 3,129 | 16,392 | 12,839 | 1,928 |
| | (Rate per $10^5$) | (0) | (0) | (0) | (12.20) | (77.89) | (103.73) |
| 55–64 | Deaths | 25 | 6 | 31 | 183 | 245 | 63 |
| | Person-years | 213,858 | 14,624 | 45,217 | 151,664 | 103,020 | 19,649 |
| | (Rate per $10^5$) | (11.69) | (41.03) | (68.56) | (120.66) | (237.82) | (320.63) |
| 65–74 | Deaths | 49 | 10 | 44 | 239 | 194 | 50 |
| | Person-years | 171,211 | 10,053 | 37,130 | 101,731 | 50,045 | 8,937 |
| | (Rate per $10^5$) | (28.62) | (99.47) | (118.50) | (234.93) | (387.65) | (559.47) |
| 75+ | Deaths | 4 | 1 | 5 | 15 | 7 | 3 |
| | Person-years | 8,489 | 512 | 1,923 | 3,867 | 1,273 | 232 |
| | (Rate per $10^5$) | (47.12) | (195.31) | (260.01) | (387.89) | (549.88) | (1,293.10) |

* Adapted from Kahn (15).

as a possible confounder) for these data can be obtained by the Poisson regression product model specified in equation 1, by assuming that $\alpha_j = \alpha$ for $j = 1, \ldots, J$, such that $\lambda_{jk} = \lambda\phi_k = \exp(\alpha + \delta_k)$, where the smoking categories are indexed by $k = 1, \ldots, K$. If nonsmokers are considered to be the reference group, then $\delta_1 = 0$ and $\lambda_{jk} = \exp(\beta_k)$, where $\beta_k = \alpha + \delta_k$ $(k = 2, \ldots, K)$. The maximum likelihood estimates of the $\beta_k$ are $\hat{\beta}_k = \ln(y_{.k}/c_{.k})$, where $y_{.k} = \sum_j y_{jk}$ and $c_{.k} = \sum_j c_{jk}$. The crude risk ratio (CRR) for the $k$th level of the risk factor is then

$$\mathrm{CRR}_k = \exp(\delta_k) = y_{.k}/[c_{.k}(y_{.1}/c_{.1})]. \quad (5)$$

Equation 5 corresponds to equation 1 of Miettinen (6). The deviance for the model representing the crude risk ratio for smoking is 589.7 with 24 df (table 5), indicating that the variation of risk in these data cannot be explained by the risk factor alone.

Proceeding under the assumption that the risk factor, smoking, is not important, i.e., all $\delta_k = 0$, equation 1 becomes $\lambda_{jk} = \lambda_j = \exp(\alpha_j)$, and the maximum likelihood estimate of $\hat{\lambda}_j = y_{j.}/c_{j.}$. Under this model, the expected number of deaths for the $k$th risk group is $\sum_j c_{jk}\hat{\lambda}_j$, and the ratio

TABLE 5
*Poisson ANOVA table for lung cancer mortality data in table 4*

| Model | $\ln(\lambda)$ | No. of parameters | Deviance* | df |
|---|---|---|---|---|
| Minimal | $\alpha$ | 1 | 1,438.0 | 29 |
| Smoking effect | $\alpha + \delta_k$ | 6 | 589.7 | 24 |
| Age effect | $\alpha_j$ | 5 | 1,037.0 | 25 |
| Age and smoking | $\alpha_j + \delta_k$ | 10 | 12.5 | 20 |
| Complete | $\delta_{jk}$ | 30 | 0 | 0 |

* The deviance provides an absolute measure of residual (i.e., unexplained) variation and is asymptotically distributed as a chi-square random variable (see Appendix).

$$y_{.k} / \sum_j c_{jk}\hat{\lambda}_j = \mathrm{SMR}_k \quad (6)$$

is an estimate of the standardized mortality ratio (SMR) for risk group $k$, where the reference group is all risk groups combined. Breslow and Day (10) have shown that the ratios, $\mathrm{SMR}_k/\mathrm{SMR}_1$ (group 1 = nonsmokers in this example), provide estimates

TABLE 6
*Comparison of summary relative risk estimates for lung cancer data in table 4*

| Relative risk estimator | Cigarettes per day | | | | | |
|---|---|---|---|---|---|---|
| | Nonsmokers* | Occasional | 1–9 | 10–20 | 21–39 | 40+ |
| Crude risk ratio | 1 | 3.2 | 4.8 | 7.5 | 12.6 | 19.3 |
| $SMR_k/SMR_1$† | 1 | 3.4 | 4.8 | 8.8 | 15.7 | 21.9 |
| $\hat{\phi}$ from product model‡ | 1 | 3.5 | 4.8 | 8.9 | 16.2 | 22.6 |

* Reference category.

† All groups combined as the standard (see equation 6). SMR, standardized mortality ratio.

‡ From Poisson regression model $\lambda_{jk} = \exp(\alpha_j + \delta_k)$, where $\delta_k$ is the natural logarithm of the standardized rate ratio, i.e., $\hat{\delta}_k = \ln\hat{\phi}_k$.

of the summary relative risks. These estimates may be confounded, however, in that they lack a common standard. The values of these ratios for the data in table 4 are given in table 6. The deviance for this model is 1,037.0 with 25 df (see table 5), thus indicating that the confounder, age, alone cannot account fully for the variability in the data.

From the foregoing, it is apparent that neither smoking nor age alone can explain adequately the variation in these data. A further analytic procedure is to fit the product model equation (equation 1) which includes both of these factors. The maximum likelihood estimates for this model cannot be expressed in closed form; therefore, the iteratively reweighted least squares procedure is used to obtain maximum likelihood estimates of the parameters (see Appendix). The deviance for this model is 12.5 with 20 df (table 5), indicating that the product model provides a very good description of these data.

Miettinen (6) has proposed using the standardized risk ratio (SRR) as a summary relative risk estimator. The standardized risk ratio for risk group $k$ is defined as

$$SRR_k = \left(\sum_j w_j y_{jk}/c_{jk}\right) \Big/ \left(\sum_j w_j \lambda_j\right), \qquad (7)$$

where the $w_j$ are the standard population weights. If the nonexposed risk group ($k = 1$) is considered to be the reference and the standard, $w_j = c_{j1}$ and $\lambda_j = y_{j1}/c_{j1}$. When the product model provides a good fit, as indicated by the contrast of the deviance and its degrees of freedom, the $y_{jk}$ in equation 7 can be replaced by their predicted values under the product model, i.e., $\hat{y}_{jk} = c_{jk}\exp(\hat{\alpha}_j + \hat{\delta}_k) = c_{jk}\hat{\lambda}_j\hat{\phi}_k$. Thus, the $SRR_k$ can be expressed as

$$SRR_k^* = \left(\sum_j w_j \hat{\lambda}_j \hat{\phi}_k\right) \Big/ \left(\sum_j w_j \hat{\lambda}_j\right) = \hat{\phi}_k = \exp(\hat{\delta}_k). \qquad (8)$$

The $\hat{\phi}_k$ are therefore estimators of the standardized risk ratios with the nonexposed group as the referent, and the choice of the standard population weights ($w_j$) is unimportant. In these situations, we recommend use of the log-linear parameterized form of the product model, as given in equation 1, because of ease of implementation with widely available statistical packages, such as GLIM (13).

If the product model provides a reasonable fit to the data, one can proceed to test the hypothesis of primary interest, i.e., that the $SRR_k = 1$ for all $k$ risk groups, against the general alternative that $SRR_k \neq 1$. This is equivalent to testing the hypothesis that $\delta_2 = \delta_3 = \cdots = \delta_k = 0$ in the product model. In the example under consideration, the test statistic is obtained by subtracting 12.5 from 1,037.0 in table 5 to obtain 1,024.5 with 5 df, indicating that differences in standardized risk ratios are highly significant.

### *Example 3—Modeling disease rates according to levels of the exposure variable and covariate*

As mentioned previously, summarizing relative risk across strata of a covariate is

appropriate when there is no interaction between relative risk and levels of the covariates (7). In the context of the product model given in equation 1, interaction is equivalent to nonadditivity on a logarithmic scale, and the deviance is used to measure the lack of fit of the product model. In some situations, an additive model $\lambda_{jk} = \lambda_j + \phi_k$ ($\phi_1 = 0$) may be more appropriate. Note that the model matrix for this additive model is identical with that for the product model, and Poisson regression can be used to fit either model. When there are more complex relationships between the rates and the covariates, conventional summary estimation techniques may not be appropriate. An alternative approach to this problem is offered by Poisson regression which can accommodate nonlinear modeling of disease rates. An example is given below to illustrate the approach.

Doll (16), in studying the association between cigarette smoking and lung cancer among British physicians, proposed a model in which the age-specific death rate is proportional to smoking rate and age, i.e.,

$$\lambda_{jk} = (\gamma + \alpha d_k^{\theta}) t_j^{\beta}, \qquad (9)$$

where $t$ = (age − 20 years)/42.5 and $d$ = exposure rate, expressed as cigarettes per day. Frome (9) has provided a detailed presentation of this model which is intrinsically nonlinear in the unknown parameter. The parameter $\gamma$ represents the lung cancer death (per $10^5$ man-years) in nonsmokers ($d = 0$) at age = 62.5 ($t = 1$), and $\gamma t^{\beta}$ corresponds to the age-specific death rate in nonsmokers at age $t$. A plot of the death rates against $t$ on a log-log scale will result in a straight line with slope of $\beta$ and intercept $\gamma$ (see example 1). The effect of smoking is represented by the term $\alpha d^{\theta}$ in equation 9, and when multiplied by $t^{\beta}$ corresponds to the increase in the lung cancer death rate for individuals that smoke $d$ cigarettes per day. Note that if $\theta = 1$, the relative *increase* in the age-specific death rate is proportional to the exposure rate for all ages, i.e., the rate ratio is $1 + (\alpha/\gamma)d$. If $\theta \neq 1$, this exposure-effect relationship will be concave ($\theta < 1$) or convex ($\theta > 1$) toward the exposure axis. The lung cancer incidence data for smokers and nonsmokers (16, 17) are shown in table 7. Poisson regression was used to fit the model specified in equation 9 for these data and, for comparison, for the lung cancer and smoking data from Kahn (15) analyzed in the previous example. The parameter estimates obtained from fitting this model for both data sets are shown in table 8. The deviance for the data of Doll (16) and the data of Doll and Hill (17) is 59.6 with 59 df, indicating a good fit, while the deviance for Kahn's data, 43.5 with 26 df, suggests a considerably poorer fit. Of particular concern with the data from Kahn is the estimate for $\theta$ which is less than 1.0, indicating a concave dose-effect relationship. The different appearances of the dose-effect curves for these two data sets are depicted in figure 2, which shows the maximum likelihood estimates of age-adjusted relative risk, according to level of smoking, obtained under the product model.

## Discussion

In this paper, we have demonstrated the use of Poisson regression modeling for summarizing relative risk estimates. The attractive features of the Poisson regression approach are that reasonable summary estimates of relative risk can be obtained, an evaluation of the presence and nature of interaction of relative risk with covariates is part of the analysis, and modeling of disease rates is facilitated. By contrast, conventional rate standardization techniques familiar to epidemiologists are most useful under the assumption of no interaction with the covariate(s), and the examination for interaction typically is limited to testing for departures from linearity. Complex forms of interaction, such as nonlinear relationships, are accommodated better by Poisson regression.

The results obtained from Poisson

TABLE 7

*Person-years at risk and number of lung cancer cases (in parentheses) from study of British physicians* according to level of cigarette consumption*

| Age minus 20 years | Cigarettes per day (median value) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0 (0) | 1–9 (5.2) | 10–14 (11.2) | 15–19 (15.9) | 20–24 (20.4) | 25–34 (27.4) | 35+ (40.8) |
| 15–19 | 10,366 (1) | 3,121 | 3,577 | 4,317 | 5,683 | 3,042 | 670 |
| 20–24 | 8,162 | 2,937 | 3,286 (1) | 4,214 | 6,385 (1) | 4,050 (1) | 1,166 |
| 25–29 | 5,969 | 2,288 | 2,546 (1) | 3,185 | 5,483 (1) | 4,290 (4) | 1,482 |
| 30–34 | 4,496 | 2,015 | 2,219 (2) | 2,560 (4) | 4,687 (6) | 4,268 (9) | 1,580 (4) |
| 35–39 | 3,512 | 1,648 (1) | 1,826 | 1,893 | 3,646 (5) | 3,529 (9) | 1,336 (6) |
| 40–44 | 2,201 | 1,310 (2) | 1,386 (1) | 1,334 | 2,411 (12) | 2,424 (11) | 924 (10) |
| 45–49 | 1,421 | 9,270 | 988 (2) | 849 | 1,567 (9) | 1,409 (10) | 556 (7) |
| 50–54 | 1,121 | 710 (3) | 684 (4) | 470 | 857 (7) | 663 (5) | 255 (4) |
| 55–59 | 826 (2) | 606 | 949 (3) | 280 | 416 (7) | 284 (3) | 104 (1) |

* Data from Doll (16) and Doll and Hill (17).

TABLE 8

*Maximum likelihood estimates for parameters specified by a nonlinear model* for lung cancer and cigarette smoking data*

| Parameter | Data source | |
|---|---|---|
| | Kahn† | Doll and Hill‡ |
| $\beta$ | 3.38 (0.18)§ | 4.46 (0.33) |
| $\log \alpha$ | 2.62 (0.21) | 1.82 (0.66) |
| $\theta$ | 0.83 (0.06) | 1.29 (0.20) |
| $\log \gamma$ | 2.61 (0.13) | 2.94 (0.58) |
| Deviance | 43.5 | 59.6 |
| df | 26 | 59 |
| $p$ value | 0.017 | 0.45 |

* Death rate = $\lambda_{jk} = (\gamma + \alpha d_k^{\theta})t_j^{\beta}$, where $t$ = (age − 20 years)/42.5 and $d$ = cigarettes per day.

† Data from Kahn (15).

‡ Data from Doll (16) and Doll and Hill (17).

§ Standard deviation in parentheses.

regression analysis are summarized in a Poisson ANOVA table which gives the deviances as measures of residual variation for the model parameters. A general $\chi^2$ test statistic for the null hypothesis of "no effect" of the risk factor, after adjustment for the effect of the covariate(s), can also be obtained from the ANOVA table (9). When this test is significant, and there are quantitative values associated with levels of the exposure variable, the covariate, or both, regression models can be developed to describe more precisely the relationships between disease rates and the study factors.

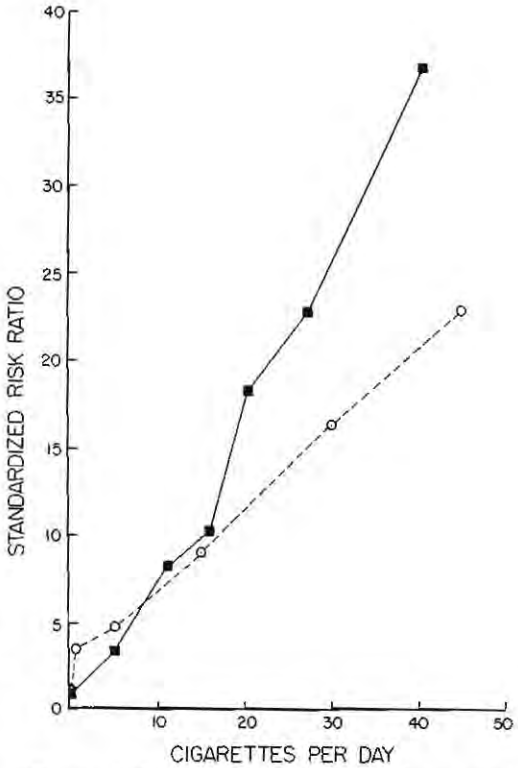


FIGURE 2. Standardized risk ratios for lung cancer and level of cigarette smoking from studies of US veterans (O- - -O) (15) and British physicians (■- - -■) (16, 17).

The flexibility of modeling offered by Poisson regression may facilitate comparisons of results from different studies when, for example, the levels of the exposure varia-

bles differ between studies. This situation was illustrated in example 3 in which an intrinsically nonlinear model was used to describe the effects of cigarette smoking on the age-specific mortality rates for lung cancer. Data from two studies that used different exposure grouping schemes for age and smoking intensity were evaluated with the same regression model. Although the examples considered here were limited to one risk factor and one covariate, the methods described can be extended to include multiple risk factors and covariates simultaneously.

The computational requirements for Poisson regression are sufficiently complex that, in most situations, a computer-based analysis would be required. High quality, inexpensive portable programs, such as GLIM (13), are now widely available, and can be used for all the analyses discussed in this paper (see Appendix for details). Poisson regression analysis can also be performed on any micro (personal) computer with software that supports ANSI standard FORTRAN using the special purpose program written by Frome (18). Consequently, while the computations required for these methods are extensive, by comparison with desk calculator or package program routines, the computational complexities should no longer limit the availability and usefulness of Poisson regression analysis.

Poisson regression models are especially appropriate in follow-up studies in which time-based denominators (person-years) are used to obtain disease rates, or when the outcome of interest is rare, such that the Poisson approximation to the binomial can be used (see Gart (19) for a discussion of this type of application). Further consideration of the use of generalized linear models for covariance adjustment and standardization is offered by Lane and Nelder (20) and Little and Pullum (21). Holford (11) has provided an excellent review of multiplicative models for rates and methods for analyzing categoric and censored survival data.

The use of Poisson regression in cohorts, with internal standard populations, has been illustrated by Frome and Hudson (22) and by Lushbaugh et al. (23) in occupational studies of radiation-exposed workers. Breslow et al. (24) have considered both internal and external standard populations in their demonstrations of the use of Poisson regression analysis and have discussed the Poisson regression approach relative to other statistical methods used in cohort studies. When an external standard set of population rates is used, e.g., national rates, the $c_{jk}$ (see table 1) are "expected" events rather than person-years, and, although the $c_{jk}$ are in fact random, the log-likelihood given under the Poisson assumption is still appropriate (24). Breslow (25) further discussed the use of Poisson regression with external rates and has illustrated how this multivariate approach can be used in relationship to traditional methods of cohort data analysis.

The more general, intrinsically nonlinear models are also of interest and can be handled readily using the iteratively reweighted least squares procedure as was illustrated in example 3. James and Segal (26) have also described the fitting of intrinsically nonlinear models of age-year interaction effects to Poisson-distributed data. Further discussion concerning the mathematical basis for regression methods appropriate to Poisson and binomial data is given by Nelder and Wedderburn (27), and Charnes et al. (28) have detailed the underlying assumptions and applications of iteratively reweighted least squares model fitting for general regression models. Frome (29) has recently reviewed the use of binomial and Poisson regression models in biomedical studies, and a general overview of Poisson regression and its relationship to other estimation procedures has been presented by Koch et al. (30).

Another useful feature of Poisson regression analysis is the availability of regression diagnostics that can be used to aid the analyst in detecting the outlying data

points or inadequacies in the model specification. Pregibon (31) has described the essential elements of regression diagnostics for logistic regression, and Frome (9) has discussed regression diagnostics for Poisson-distributed data.

## References


1. Breslow NE, Day NE. Fundamental measures of disease occurrence and association. Chap 2. In: Statistical methods in cancer research: analysis of case-control studies. Vol 1. Lyon: IARC, 1980.
2. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. JNCI 1959;22:719–48.
3. Rothman KJ, Boice JD. Epidemiologic analysis with a programmable calculator. Washington, DC. NIH publication no. 79-1649, 1979:12–13.
4. Tarone RE. On summary estimators of relative risk. J Chronic Dis 1981;34:463–8.
5. Lilienfeld DE, Pyne DA. On indices of mortality: deficiencies, validity and alternatives. J Chronic Dis 1979;32:463–8.
6. Miettinen OS. Standardization of risk ratios. Am J Epidemiol 1972;96:383–8.
7. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982:320–63.
8. Frome EL, Kutner MH, Beauchamp JJ. Regression analysis of Poisson-distributed data. J Am Stat Assoc 1973;68:935–40.
9. Frome EL. The analysis of rates using Poisson regression models. Biometrics 1983;39:665–74.
10. Breslow NE, Day NE. Indirect standardization and multivariate models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. J Chronic Dis 1975; 28:289–303.
11. Holford TR. The analysis of rates and survivorship using log-linear models. Biometrics 1980; 36:299–305.
12. Scotto J, Kopf AW, Urbach F. Non-melanoma skin cancer among Caucasians in four areas of the United States. Cancer 1974;34:1333–8.
13. Baker RJ, Nelder JA. Generalized linear interactive modeling (GLIM). Release 3, Numerical Algorithms Group, Oxford, England, 1978.
14. Armitage P, Doll R. Stochastic models for carcinogenesis. In: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability: Biology and problems of health. Vol 4. Berkeley, CA: University of California Press, 1961:19–38.
15. Kahn HA. The Dorn study of smoking and mortality among US veterans: report on eight and one-half years of observation. In: Haenszel W, ed. Epidemiological approaches to the study of cancer and other chronic diseases. Natl Cancer Inst Monogr 19, 1966:1–125.
16. Doll R. The age distribution of cancer: implications for models of carcinogenesis. J R Stat Soc (A) 1971;135:133–66.
17. Doll R, Hill AB. Mortality of British doctors in relation to smoking: observations on coronary thrombosis. In: Haenszel W, ed. Epidemiological approaches to the study of cancer and other chronic diseases. Natl Cancer Inst Monogr 19, 1966:205–58.
18. Frome EL. Poisson regression analysis. Am Statistician 1981;35:262–3.
19. Gart JJ. The analysis of ratios and cross-product ratios of Poisson variates with application to incidence rates. Communications Statist Theor Meth 1978;A7(10):917–37.
20. Lane PW, Nelder JA. Analysis of covariance and standardization as instances of prediction. Biometrics 1982;38:613–21.
21. Little RJA, Pullum TW. The general linear model and direct standardization. Sociol Meth Res 1979;7:475–501.
22. Frome EL, Hudson DR. A general statistical data structure for the epidemiologic studies of DOE workers. In: Truett T, Margolies D, Mensing RW, eds. Proceedings of the 1980 Department of Energy Statistical Symposium, CONF-801045, Berkeley, CA, 1981:206–18.
23. Lushbaugh CC, Fry SA, Shy CM, et al. DOE health and mortality study at Oak Ridge. In: Proceedings of the Health Physics Society, sixteenth midyear topical meeting, CONF-83101, Albuquerque, NM, 1983:105–14.
24. Breslow NE, Lubin JH, Marek P, et al. Multiplicative models and cohort analysis. J Am Stat Assoc 1983;78:1–12.
25. Breslow N. Multivariate cohort analysis. JNCI (in press).
26. James IR, Segal MR. On a method of mortality analysis incorporating age-year interaction, with application to prostate cancer mortality. Biometrics 1982;38:433–43.
27. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc (A) 1972;135:370–84.
28. Charnes A, Frome EL, Yu PL. The equivalence of generalized least squares and maximum likelihood estimation in the exponential family. J Am Stat Assoc 1976;71:169–72.
29. Frome EL. Regression methods for binomial and Poisson-distributed data. In: Herbert D, ed. Proceedings of the American Association of Physicists in Medicine, first midyear topical symposium. Multiple regression analysis: application in the health sciences (in press).
30. Koch GG, Atkinson SS, Stokes ME. Poisson regression. In: Kotz S, Johnson LL, Read A, eds. Encyclopedia of statistical sciences. Vol 6. New York: John Wiley & Sons (in press).
31. Pregibon D. Logistic regression diagnostics. Ann Stat 1981;9:705–24.
32. McCullagh P. Quasi-likelihood functions. Ann Stat 1983;11:59–67.
33. Goodnight JH, Sale JP. SAS user's guide: statistics. Cary, NC: SAS Institute, 1982.
34. Biomedical Computer Programs P-Series (BMDP-79). Berkeley, CA: University of California Press, 1979.

## APPENDIX

*Illustration of the model specification and computations involved in Poisson regression analysis*

The product model (see equation 1 and table 1) can be expressed as a log-linear model as follows:

$$\lambda_{jk} = \exp(\alpha_j + \delta_k) = \exp(X_i\boldsymbol{\beta}),$$

where $X_i = (x_{i1}, \ldots, x_{ip})$ is a row vector of indicator variables for the $i$th cell in the table and $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_j, \delta_2, \ldots, \delta_K)'$ is the $p$-dimensional column vector of unknown parameters with $p = J + K - 1$. If the $i$th cell of the table corresponds to row $j$ and column $k$, the components of $X_i$ ($i = 1, \ldots, JK$) can be defined as follows:

$$x_{im} = 1 \quad \text{if} \quad m = j;$$

$$x_{im} = 1 \quad \text{if} \quad k > 1 \text{ and } m = J + k - 1$$

$$(m = 1, \ldots, J + K - 1);$$

$$x_{im} = 0 \text{ otherwise.}$$

The rows of the $JK$ by $J + K - 1$ $\boldsymbol{X}$ matrix for a $J = 4$ by $K = 3$ table are shown in Appendix table 1 to illustrate the general procedure.

When $c_{jk} > 0$ for all $j$ and $k$, this situation is equivalent to a full rank parameterization of the design matrix for a two-factor fixed effects ANOVA model, and the $\delta_k$ are the natural logarithms of the standardized rate ratios with $k = 1$ as the standard and reference group (see example 2). In practice, it is not necessary to generate this matrix because its structure is implied by the levels of the factors.

Frome (9) has shown that maximum likelihood estimates of the parameter vector $\boldsymbol{\beta}$ is equivalent to a weighted least squares regression. Consider the following weighted sum of squares:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} w_i[r_i - \lambda(X_i, \boldsymbol{\beta})]^2,$$

where $r_i = y_i/c_i$ denotes the observed rate in the $i$th cell ($i = 1, \ldots, N$), where $N$ is the number of cells in the table, and the "Poisson weights" are $w_i = c_i/\lambda(X_i, \boldsymbol{\beta})$. The maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ is obtained using the iteratively reweighted least squares procedure, and the deviance is computed to obtain an absolute measure of residual (unexplained) variation. The deviance is a likelihood ratio statistic—$D(\hat{\boldsymbol{\beta}}) = -2[L(\hat{\boldsymbol{\beta}}) - L(\boldsymbol{y})]$—where $L(\hat{\boldsymbol{\beta}})$ denotes the log-likelihood function evaluated at the maximum likelihood estimate, and $L(\boldsymbol{y})$ is the log-likelihood function for the saturated model $\hat{\mu}_i = y_i$ ($i = 1, \ldots, N$).

The deviance (see references 9 and 27) for a Poisson regression model is defined as follows:

$$D(\hat{\boldsymbol{\beta}}) = 2 \sum_{i=1}^{N} [y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)],$$

where the fitted value $\hat{\mu}_i = c_i\lambda(X_i, \hat{\boldsymbol{\beta}})$, and

APPENDIX TABLE 1
*Rows of the $\boldsymbol{X}$ matrix for a 4 by 3 table*

| $j$ | $k$ | $i$ | $x_{i1}$ | $x_{i2}$ | $x_{i3}$ | $x_{i4}$ | $x_{i5}$ | $x_{i6}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 2 | 2 | 1 | 0 | 0 | 0 | 1 | 0 |
| 1 | 3 | 3 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 4 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 2 | 5 | 0 | 1 | 0 | 0 | 1 | 0 |
| 2 | 3 | 6 | 0 | 1 | 0 | 0 | 0 | 1 |
| 3 | 1 | 7 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 | 2 | 8 | 0 | 0 | 1 | 0 | 1 | 0 |
| 3 | 3 | 9 | 0 | 0 | 1 | 0 | 0 | 1 |
| 4 | 1 | 10 | 0 | 0 | 0 | 1 | 0 | 0 |
| 4 | 2 | 11 | 0 | 0 | 0 | 1 | 1 | 0 |
| 4 | 3 | 12 | 0 | 0 | 0 | 1 | 0 | 1 |

```
$SUBFILE DATA !  FCAPP1.RSS  ON 21 FEB 83
!                    FCAPP1:  APPENDIX  !
!    LUNG CANCER MORTALITY-- AGE BY CIGARETTE CONSUMPTION !
!    SOURCE: KAHN(1966) APPENDIX TABLE A !
!                    COLUMN
!     1      2      3      4      5      6
!    NEVER  OCCA   1-9   10-20  21-39  40+
$UNITS 30 !
$DATA C  $READ                 !  C =  MAN YEARS
    35164   3657   8063   59965  40643  3992
    15134  1283    3129   16392  12839  1928
   213858  14624  45217  151664 103020  19649
   171211  10053  37130  101731  50045  8937
     8489   512    1923    3867   1273   232
$DATA Y  $READ                 ! Y = LUNG CANCER DEATHS
    0   0   0    2    4   0
    0   0   0    2   10   2
   25   6  31  183  245  63
   49  10  44  239  194  50
    4   1   5   15    7   3
$M TITLE  FCAPP1: LUNG CANCER MORTALITY (KAHN.1966) $E
$CA %R=5 : %L=6  $FAC  ROW %R COL %L !
$CA ROW=%GL(%R,%L)  :  COL=%GL(%L,1) !  GENERATE ROW AND COL
$DATA 6 DOSE $READ 0 0.5 5 15 30 45  !  CIG. PER DAY
$CA D=DOSE(COL) : AG= ROW*10+30      !
!        DEFINE VARIABLES X1 AND X2 FOR NONLINEAR MODEL !
$WARN $CA X1= %IF( %LE(D,0),-8,%LOG(D) ) !
$CA  X2=%LOG( (AG-20)/42.5) !
$CA C=C/100000 : R=Y/C !
$PR TITLE : 'R = LUNG CA DEATHS / 100000 MAN-YEARS'$
$RETURN
```

FIGURE 1A. GLIM (13) program statements for defining data structure for Poisson regression analysis using the product model for data from Kahn (15) in example 2.

$\lambda(X_i, \hat{\beta})$ denotes the regression function evaluated at the maximum likelihood estimate $\hat{\beta}$. (Note that maximizing the log-likelihood function is equivalent to minimizing the deviance.) Another well known statistic that can be used as a measure of residual variation in Poisson regression is Pearson's $X^2 = \sum_i (y_i - \hat{\mu}_i)^2/\hat{\mu}_i$ (see reference 7). Both of these test statistics will yield similar values when the fitted values are large (i.e., $\hat{\mu}_i > 3$ for all $i = 1, \ldots, N$). The value of these statistics may differ substantially when some of the $\hat{\mu}_i$ are small, and the analyst should proceed with caution. If the assumed regression function $\lambda(X, \beta)$ is appropriate, both the deviance and Pearson's $X^2$ will be distributed approximately as a chi-square with $N - p$ df, where $N$ is the number of observations and $p$ is the number of parameters (see references 8 and 32). It is generally advisable to utilize regression diagnostics (9, 31) to assess the effect of outlying $y$ values and/or model inadequacies on these lack of fit statistics and the estimated parameter values. In situations in which the unexplained variation is substantial, the lack of fit may be due to overdispersion (relative to the Poisson distribution) or to the inadequacy of the regression function. If the lack of fit is attributed to overdispersion, the estimated parameter covariance matrix should be multiplied by the heterogeneity factor $\hat{\sigma}^2 = X^2/(N - p)$, approximate interval estimates, for the components of $\beta$ should be based on the $t$ distribution, and test

```
$SUBFILE PMOD  !
$MAC PMOD        !  FIT PRODUCT MODEL
$ERR P  $WE C  $YVAR R $FIT !
$VAR %L BZ II $CA II=%CU(1) !
$FIT COL $DISP E $USE RR     !
$FIT +ROW $DISP E $USE RR $E !
$MAC RR          ! CALCULATE RISK RATIOS FOR PMOD
 $EXTR %PE $CA  BZ(II)=%PE(II) : BZ(1)=0 : BZ=%EXP(BZ) !
$PR : '   RISK  RATIOS = ' BZ : $END
$RETURN
!
$SUBFILE NONLIN
$MAC FITNL       ! FIT NONLINEAR MODEL DEFINED BY MAC NLM
$DATA 4 B $READ 4.5 2 1 2 ! STARTING VALUES FOR BETA
$CA %K=20 : %C= 0.00001    ! SET CONCERGENCE CRITERIA
$WE W  $OWN R1 R2 R3 R4 $SC 1 $YVAR Z ! DEFINE MODEL
$WHILE %K NLM !
$PR : '                  STAND DEVS ' $DISP E $ !
$EXTR %VL $CA H= %VL*W  !
$DEL V1 V2 P1 P2 P3 P4 DB Z    ! DELETE WORK ARRAYS
$PR  'H CONTAINS DIAG TERMS FROM H MATRIX' !
$CA FT=%SQRT(Y)+%SQRT(Y+1)-%SQRT(4*C*RHAT+1) !
$PR '  FT CONTAINS  FREEMAN-TUKEY RESIDUALS' !
$DEL NLM R1 R2 R3  !
$END
$MACRO NLM      !
!          POISSON RATE ANALYSIS !
!  MACROS TO FIT NONLINEAR MODEL USING IRLS !
$CA V1= %EXP(B(1)*X2)!
$CA V2= %IF( %LE(D,0.1),0.0,%EXP(B(2)+B(3)*X1) ) !
$CA RHAT= V1*( V2+ %EXP(B(4)) ) $CA P1= RHAT*X2 !
$CA P2= V2*V1 $CA P3= P2*X1 $CA P4= V1*%EXP(B(4)) !
$CA W= C/RHAT : Z= R-RHAT : %LP= Z    !
$FIT P1+P2+P3+P4-%GM $EXTR %PE  $CA DB=%PE !
$CA B= B + DB $PR %K  '  ESTIMATES=' B  !
!      CHECK FOR CONVERGENCE OR MAXIMUM ITERATIONS !
$CA DB=DB/B :  DB= %IF( %LE(DB,0),-DB,DB) !
$CA DB= %IF( %LE(DB,%C),0,1)  $CA %T= %CU(DB)!
$CA %K=%K-1 $CA %K= %IF( %LE(%T,0),0,%K) $ !
$END
!    MACROS REQUIRED BY OWN FOR POISSON RATE ANALYSIS !
$M R1 $CA %FV=%LP$E $M R2 $CA %DR=1.0 $E $M R3 $CA %VA=1.0$E !
$M R4 $CA %DI= 2*( Y*%LOG(R/RHAT)-C*Z)/W $E !
$RETURN
```

FIGURE 2A. GLIM (13) program statements for Poisson regression analysis of Kahn's (15) data using the product model (example 2) and the nonlinear model in equation 9 (example 3).

statistics for the relative importance of subsets of parameters should be based on approximate $F$ ratios (see references 8 and 32).

Figure 1A provides a listing of the GLIM (13) program statements that were used for the analysis of the data in example 2 from Kahn (15). The GLIM statements in figure 1A read the number of person-years into the vector $C$ and the corresponding number of lung cancer deaths into the vector $Y$. In GLIM terminology, ROW and COL are factors that correspond to age and smoking, respectively. The values of these factors and the variables $X1$ and $X2$ that are needed for the nonlinear model are generated in the last 10 lines of figure 1A. The GLIM macro PMOD in figure 2A was used for the analysis of the data from example 2 using the product model.

The data from example 3 necessitated use of a model which is nonlinear in its parameter (equation 9) and therefore cannot be fitted to these data with standard

GLIM options. Maximum likelihood estimates of the model parameters can be obtained from the iteratively reweighted linear squares algorithm using a GLIM macro developed for this purpose (9). The GLIM macro statements for fitting this nonlinear model are shown in figure 2A. The iteratively reweighted linear squares estimation method requires the partial derivatives of the rate function

$$\lambda(X_i, \boldsymbol{\beta}) = [\exp(\beta_2 + \beta_3 x_{i1}) + \exp(\beta_4)] \exp(\beta_1 x_{i2}),$$

where $x_{i1} = \ln(d_i)$ ($d_i$ = dose rate), $x_{i2} = \log(t_i)$, and $\boldsymbol{\beta} = (\beta, \ln\alpha, \theta, \ln\gamma)'$. The partial derivatives are defined in the GLIM macro NLM (see figure 2A) as $P1$, $P2$, $P3$, and $P4$, where, for example, $P1$ is

$$P1 = \partial\lambda(X, \boldsymbol{\beta})/\partial\beta_1 = x_{i2}[V2 + \exp(\beta_4)]V1,$$

where $V2 = \exp(\beta_2 + \beta_3 x_{i1})$ and $V1 = \exp(\beta_1 x_{i2})$. Additional macros that are required for the nonlinear model are also listed in figure 2A, and the reader is referred to the GLIM manual (13, Chap. 18) for further details. Identical results can also be obtained using the FORTRAN program PREG (18), the SAS procedure NLIN (33), and the BMDP program P3R (34).