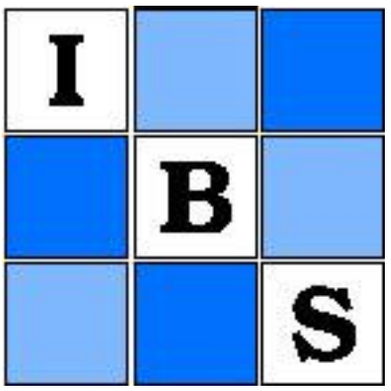



BIOMETRICS 39, 665–674
September 1983

# The Analysis of Rates Using Poisson Regression Models

E. L. Frome[1]

Medical and Health Sciences Division, Oak Ridge Associated Universities,
Oak Ridge, Tennessee 37830, U.S.A.

SUMMARY

Models are considered in which the underlying rate at which events occur can be represented by a regression function that describes the relation between the predictor variables and the unknown parameters. Estimates of the parameters can be obtained by means of iteratively reweighted least squares (IRLS). When the events of interest follow the Poisson distribution, the IRLS algorithm is equivalent to using the method of scoring to obtain maximum likelihood (ML) estimates. The general Poisson regression models include log-linear, quasilinear and intrinsically nonlinear models. The approach considered enables one to concentrate on describing the relation between the dependent variable and the predictor variables through the regression model. Standard statistical packages that support IRLS can then be used to obtain ML estimates, their asymptotic covariance matrix, and diagnostic measures that can be used to aid the analyst in detecting outlying responses and extreme points in the model space. Applications of these methods to epidemiologic follow-up studies with the data organized into a life-table type of format are discussed. The method is illustrated by using a nonlinear model, derived from the multistage theory of carcinogenesis, to analyze lung cancer death rates among British physicians who were regular cigarette smokers.

## 1. Introduction

Data are often obtained in medical and epidemiologic studies in which the dependent variable is a count (e.g. number of cancer deaths) obtained in each of a number of subgroups that are described by a set of predictor variables. Let $y_i$ denote the number of failures and $c_i$ the total follow-up time for Subgroup $i$, $i = 1, \ldots, n$. The expected number of failures in the $i$th subgroup is

$$\mu_i = c_i \lambda(\mathbf{X}_i, \boldsymbol{\beta}), \tag{1.1}$$

where $\mathbf{X}_i = (x_{i1}, \ldots, x_{im})$ is an $m$-dimensional row vector of predictor variables that describe the $i$th subgroup, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p$-dimensional vector of unknown parameters, and $\lambda(\mathbf{X}_i, \boldsymbol{\beta})$ is the expected rate for the $i$th subgroup. The rate function $\lambda(\mathbf{X}_i, \boldsymbol{\beta})$ can be viewed as a regression function that relates the expected number of failures to the predictor variables and the parameters.

If we assume that the rate function has a log-linear form, we obtain

$$\lambda(\mathbf{X}_i, \boldsymbol{\beta}) = \exp(\mathbf{X}_i \boldsymbol{\beta}). \tag{1.2}$$

Holford (1980) has presented an excellent review of the multiplicative model for rates and has described its relationship to methods for analyzing categorical data and censored survival data. Laird and Olivier (1981) have discussed in detail the fitting of count data from survival studies by log-linear models, and the use of iterative scaling algorithms for analysis. In some situations (see, for example, Osborn, 1975), investigators have proposed that one or more

[1] Present address: Mathematics and Statistics Research Department, Computer Sciences, Oak Ridge National Laboratory, P.O. Box Y, Oak Ridge, Tennessee 37830, U.S.A.

*Key words*: Poisson regression; Nonlinear models; Iteratively reweighted least squares; Hazard rate; Log-linear models.

factors affect the failure rate in an additive way. Freeman and Holford (1980) have discussed the importance of linear and log-linear models in smoothing rates, and their relationship to the usual approaches to standardization (direct and indirect). A third possibility is a mixture of additive and multiplicative effects, and a general 'quasilinear' hazard-rate model has been proposed by Taulbee (1979). Another model that we consider in §3 is given by (3.2) and provides an example of a nonlinear regression model. Further examples of situations in the biological sciences, where the response variable is a count that follows the Poisson distribution, and the regression function is intrinsically nonlinear in the parameters, have been given by Frome and Beauchamp (1968), Hasselblad (1981) and Frome and DuFrain (1982).

A situation of special interest occurs in epidemiologic follow-up studies where data are organized into a format similar to that of a life table. One dimension of the table corresponds to the levels of one or more factors that may affect the survival experience of the cohort under study. These could include categorical variables (e.g. race, sex) or grouped values of exposure variables. The other dimension of the table is age; we let $t_j = (j - \frac{1}{2})\Delta t$ denote the midpoint of the $j$th age (or time) interval and we let $y_{jk}$ be the number of failures in Subgroup $k$ with person-years $c_{jk}$ and covariates $\mathbf{Z}_{jk}$. Let $\lambda(t, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ denote the hazard function at Time $t$, and assuming that $\lambda(t, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \lambda_0(t, \boldsymbol{\alpha})\exp(\mathbf{Z}\boldsymbol{\theta})$, we obtain Cox's proportional-hazards model, where $\lambda_0(t, \boldsymbol{\alpha})$ denotes the hazard at the standard set of conditions $\mathbf{Z} = 0$. If we let $\lambda_0(t, \boldsymbol{\alpha}) = \exp(\alpha_j)$ and consider only those values of $t_j$ where at least one failure has occurred, i.e. $y_{j\cdot} = \sum_k y_{jk} > 0$, then the rate function takes the log-linear form $\lambda(t, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \exp(\alpha_j + \mathbf{Z}_{jk}\boldsymbol{\theta})$ proposed by Cox. In particular, if $c_{jk}$ is the number of individuals 'in view' at the start of the $j$th interval with covariate values $\mathbf{Z}_{jk}$, and if there are no ties (i.e. if $y_{j\cdot} = 1$ for all $j$), then the maximum likelihood (ML) estimates obtained by using the iteratively reweighted least squares (IRLS) procedure (see §2) are equivalent to those obtained by maximizing Cox's partial likelihood. Whitehead (1980) further showed that when ties are present in the data (i.e. when $y_{jk} > 1$ for some $j$), ML estimates under the Poisson assumption yield estimates of $\boldsymbol{\theta}$ that maximize the generalized partial likelihood function based on Peto's approximation (Peto, 1972). This leads one to conclude that algorithms for fitting generalized linear models can be used to analyze censored survival data (see Holford, 1980; Whitehead, 1980; Aitkin and Clayton, 1980). In this note, we show that this type of analysis can be viewed as a weighted least squares regression and that the results can be extended to apply to any reasonable regression function—i.e. $\lambda(\mathbf{X}, \boldsymbol{\beta})$ is a differentiable function of $\boldsymbol{\beta}$.

## 2. Estimation

Let $r_i = y_i/c_i$ denote the failure rate in the $i$th subgroup and consider the following weighted sum of squares

$$S(\boldsymbol{\beta}) = \sum_i w_i\{r_i - \lambda(\mathbf{X}_i, \boldsymbol{\beta})\}^2, \tag{2.1}$$

where $w_i$ denotes a weight inversely proportional to the variance of $r_i$. Since $\lambda(\mathbf{X}, \boldsymbol{\beta})$ is, in general, nonlinear in the parameters, we replace it with the linear terms in a Taylor series expansion about an initial estimate, $\boldsymbol{\beta}^0$,

$$\lambda(\mathbf{X}_i, \boldsymbol{\beta}) \simeq \lambda(\mathbf{X}_i, \boldsymbol{\beta}^0) + P_i^0\boldsymbol{\delta}^0, \tag{2.2}$$

where $P_i^0$ denotes the $i$th row of the $n \times p$ matrix of partial derivatives $p_{ij} = \partial\lambda(\mathbf{X}_i, \boldsymbol{\beta})/\partial\beta_j$ evaluated at the initial estimate $\boldsymbol{\beta}^0$, and $\boldsymbol{\delta}^0 = (\delta_1^0, \ldots, \delta_p^0)'$. Using (2.2) in (2.1) and the least squares principle, we can obtain estimates of the $\delta_j^0$ by solving the following system of $p$ linear equations:

$$\mathbf{P}(\boldsymbol{\beta}^0)'\mathbf{W}\mathbf{P}(\boldsymbol{\beta}^0)\boldsymbol{\delta}^0 = \mathbf{P}(\boldsymbol{\beta}^0)'\mathbf{W}\{\mathbf{R} - \boldsymbol{\Lambda}(\boldsymbol{\beta}^0)\}, \tag{2.3}$$

where $\mathbf{R} = (r_1, \ldots, r_n)'$, $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$, and $\boldsymbol{\Lambda}(\boldsymbol{\beta}^0)$ denotes $\boldsymbol{\Lambda}(\boldsymbol{\beta}) = \{\lambda(\mathbf{X}_1, \boldsymbol{\beta}), \ldots, \lambda(\mathbf{X}_n, \boldsymbol{\beta})\}'$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^0$. We then obtain a revised estimate $\boldsymbol{\beta}^1 = \boldsymbol{\beta}^0 + \boldsymbol{\delta}^0$, replace the

superscripts of 0 in (2.3) with superscripts of 1, and solve for $\boldsymbol{\delta}^1$. The iterative process (Gauss–Newton method) continues until some convergence criteria are satisfied.

In many situations of interest, it is reasonable to assume that the $y_i$ are independent and follow the Poisson distribution (see, for example, Armitage, 1966; Breslow and Day, 1975; Osborn, 1975; Gail, 1978; Holford, 1980) with expectation given by (1.1). The kernel of the log likelihood function is

$$L(\boldsymbol{\beta}) = \sum_i [y_i \log\{c_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})\} - c_i \lambda(\mathbf{X}_i, \boldsymbol{\beta})], \tag{2.4}$$

and the ML equations are given by

$$\partial L(\boldsymbol{\beta})/\partial \beta_j = \sum_i p_{ij}\{y_i/\lambda(\mathbf{X}_i, \boldsymbol{\beta}) - c_i\}, \qquad j = 1, \ldots, p. \tag{2.5}$$

Since the ML equations will generally be nonlinear with respect to the unknown parameters, the method of scoring is used to develop an iterative algorithm to find a root of (2.5). This leads to the following system of equations on Iteration $k + 1$:

$$\mathbf{A}(\boldsymbol{\beta}^k)\boldsymbol{\delta}^k = \mathbf{G}(\boldsymbol{\beta}^k),$$

where $\mathbf{A}(\boldsymbol{\beta}^k)$ is the information matrix with elements

$$a_{js} = \sum_i \{(p_{ij}p_{is}c_i)/\lambda(\mathbf{X}_i, \boldsymbol{\beta})\}, \qquad j, s = 1, \ldots, p,$$

evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^k$, and $\mathbf{G}(\boldsymbol{\beta}^k)$ is (2.5) evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^k$, i.e.

$$g_j = \sum_i p_{ij}\{c_i/\lambda(\mathbf{X}_i, \boldsymbol{\beta}^k)\}\{r_i - \lambda(\mathbf{X}_i, \boldsymbol{\beta}^k)\}, \qquad j = 1, \ldots, p.$$

If the weights in (2.3) on Iteration $k + 1$ are chosen to be $w_i = c_i/\lambda(\mathbf{X}_i, \boldsymbol{\beta}^k)$ then $\mathbf{A}(\boldsymbol{\beta}^k) = \mathbf{P}(\boldsymbol{\beta}^k)'\mathbf{W}\mathbf{P}(\boldsymbol{\beta}^k)$ and $\mathbf{G}(\boldsymbol{\beta}^k) = \mathbf{P}(\boldsymbol{\beta}^k)'\mathbf{W}\{\mathbf{R} - \boldsymbol{\Lambda}(\boldsymbol{\beta}^k)\}$, i.e. the IRLS procedure is equivalent to using the method of scoring to obtain a root of the likelihood equations. This result was discussed in detail by Frome and Beauchamp (1968), Frome, Kutner and Beauchamp (1973) and E. L. Frome (in a Ph.D. Dissertation at Emory University, 1972).

This result was also reached by Nelder and Wedderburn (1972) for generalized linear models when the dependent variable is from exponential family. Extension of this result to nonlinear models, and conditions under which a solution of the likelihood equations will yield a global maximum of the likelihood function, were given by Charnes, Frome and Yu (1976). For linear and log-linear models, a solution to (2.5) will be the unique ML estimate, and a solution will exist if the columns of the $\mathbf{X}$ matrix are linearly independent when the rows with $y_i = 0$ are excluded (see Nelder and Wedderburn, 1972). An algorithm for obtaining the ML estimate of $\boldsymbol{\beta}$ that is coded in ANSI Standard FORTRAN was given by Frome, (1981, and in his 1972 dissertation). The log-linear models with Poisson 'errors' can also be fitted as a standard option in the statistical package GLIM (Baker and Nelder, 1978). It is also possible to fit nonlinear models in GLIM by using the IRLS approach to develop model-specific procedures that can be implemented by using GLIM macros.

Pregibon (1981) has proposed that when regression methods are used in observational studies, diagnostic procedures (similar to those developed for the standard linear model) should be used to check for outlying $y$-values and extreme points in the 'model' space. The basic 'building blocks' that are required for various diagnostic measures are standardized residual of some type and the diagonal terms, $h_i$, from the matrix

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{P}(\mathbf{P}'\mathbf{W}\mathbf{P})^{-1}\mathbf{P}'\mathbf{W}^{\frac{1}{2}}, \tag{2.6}$$

where all quantities that depend on $\boldsymbol{\beta}$ are evaluated at the ML estimate $\hat{\boldsymbol{\beta}}$. The diagonal terms of this matrix are useful in detecting extreme points in the model space that may have a substantial influence on the fitted model. Recall that for the standard linear model $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{I} - \mathbf{H}$ is the projection matrix, and large values of $h_i$ denote extreme points in

the model (design) space. For generalized linear models $\lambda(\mathbf{X}_i, \boldsymbol{\beta}) = g(\eta_i)$, where $\eta_i = \sum_j \beta_j x_{ij}$, and $\mathbf{H}$ can be written as

$$\mathbf{H} = \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}'\mathbf{V}^{\frac{1}{2}},$$

where $\mathbf{V}$ is diagonal with $v_i = \{c_i(\partial g_i/\partial \eta_i)^2\}/g_i$. Note that $\sum_i h_i = p$ and that large values of $h_i$ (say, greater than $2p/n$) indicate extreme points in the model space. If $u_i$ denotes a standardized residual [e.g. $(y_i - \hat{\mu}_i)/\hat{\mu}_i^{\frac{1}{2}}$], the variance of $u_i$ is approximately $1 - h_i$, and 'adjusted residuals' are given by $u_i/(1 - h_i)^{\frac{1}{2}}$ (see Haberman, 1974, Ch. 4). Another standardized residual that is useful for Poisson data is $u_i = y_i^{\frac{1}{2}} + (y_i + 1)^{\frac{1}{2}} - (4\hat{\mu}_i + 1)^{\frac{1}{2}}$ which is known as the Freeman-Tukey (FT) residual (Freeman and Tukey, 1950). A third alternative, the 'signed deviance', and various approaches to using these basic building blocks were discussed in detail by Pregibon (1981).

## 3. Example

The best dose–response data for human cancer are those obtained by Doll and Hill (1966) in a study of cigarette smoking in British physicians (see, particularly, Doll, 1971). Whittemore and Altshuler (1976) have presented a more refined analysis of Doll's data on the lung cancer mortality of cigarette smokers in several age and dose groups, and this is shown in Table 1. In addition to these data on continuing cigarette smokers whose cigarette consumption was constant, data for nonsmokers obtained from Doll and Hill (1966, Appendix, Table 5) have been added to Table 1. These data will be used to illustrate the Poisson regression methods using both log-linear and nonlinear regression models.

One approach to analyzing the lung cancer death rates in Table 1 is to use the iterative indirect standardization technique (see Mantel and Stark, 1968; Breslow and Day, 1975). This technique is equivalent to fitting a log-linear model in which the rows in the model matrix correspond to the rows in a full-rank design matrix for a two-factor fixed-effects analysis-of-variance (ANOVA) model, i.e.

$$\begin{aligned}\lambda(\mathbf{X}_i, \boldsymbol{\beta}) &= \exp(\mathbf{X}_i\boldsymbol{\beta}) \\ &= \exp(\mu + \alpha_j + \delta_k). \end{aligned} \tag{3.1}$$

Maximum likelihood estimates of the parameters are obtained by using the IRLS procedure and are used to compute estimates of the 'age fit', i.e. $\exp(\hat{\mu} + \hat{\alpha}_j)$, which are given in the last column of Table 1, and estimates of the 'smoking effects', $\exp(\hat{\delta}_k)$, which are shown at the bottom of Table 1.

It is evident from the results in Table 1 that the lung cancer death rate is related to the dose rate (cigarettes per day) and the duration of smoking. It would be possible to use the quantitative values associated with the row and column factors to develop a parsimonious description of the relation between age-specific lung cancer incidence rates and smoking rates by using empirical log-linear models. An alternative approach is to use some mathematical theory of carcinogenesis that has been proposed to describe this relation. Whittemore and Keller (1978) have pointed out that theoretical models play an important role since they provide a bridge between animal experiments and epidemiologic studies, and also because they provide a basis for extrapolating dose–response relations downward (see Crump *et al.*, 1976). One such model that has been applied to data obtained in both animal (see, for example, Carlborg, 1981) and epidemiologic studies (Whittemore and Keller, 1978) is given by

$$\lambda(t, d) = (\gamma + \alpha d^{\theta})t^{\beta}, \tag{3.2}$$

where $d$ is the amount of carcinogen applied per unit of time at a constant rate, and $t$ denotes time from the start of exposure. For the data in Table 1, $d$ is smoking rate (cigarettes per day)

**Table 1**

*Man-years at risk, number of cases of lung cancer (in parentheses), and fitted values obtained under the product model*

| Years of smoking (age minus 20 years) | Cigarettes/day: | Nonsmokers 0 | 1–9 5.2 | 10–14 11.2 | 15–19 15.9 | 20–24 20.4 | 25–34 27.4 | 35 + 40.8 | Age fit* (per 100 000 man years) |
|---|---|---|---|---|---|---|---|---|---|
| 15–19 | | 10366 (1) | 3121 | 3577 | 4317 | 5683 | 3042 | 670 | .3 |
| 20–24 | | 8162 | 2937 | 3286 (1) | 4214 | 6385 (1) | 4050 (1) | 1166 | .9 |
| 25–29 | | 5969 | 2288 | 2546 (1) | 3185 | 5483 (1) | 4290 (4) | 1482 | 1.9 |
| 30–34 | | 4496 | 2015 | 2219 (2) | 2560 (4) | 4687 (6) | 4268 (9) | 1580 (4) | 8.5 |
| 35–39 | | 3512 | 1648 (1) | 1826 | 1893 | 3646 (5) | 3529 (9) | 1336 (6) | 8.8 |
| 40–44 | | 2201 | 1310 (2) | 1386 (1) | 1334 (2) | 2411 (12) | 2424 (11) | 924 (10) | 23.2 |
| 45–49 | | 1421 | 927 | 988 (2) | 849 (2) | 1567 (9) | 1409 (10) | 556 (7) | 29.4 |
| 50–54 | | 1121 | 710 (3) | 684 (4) | 470 (2) | 857 (7) | 663 (5) | 255 (4) | 46.5 |
| 55–59 | | 826 (2) | 606 | 449 (3) | 280 (5) | 416 (7) | 284 (3) | 104 (1) | 77.3 |
| Smoking effect† | | 1.0 | 3.39 | 8.16 | 10.1 | 18.2 | 22.6 | 36.8 | |

* Age fit = $\exp(\hat{\mu} + \hat{\alpha}_j)$, where $\hat{\mu}$ and $\hat{\alpha}$ are ML estimates defined by the product model.

† Smoking effect = $\exp(\hat{\delta}_k)$, where $\hat{\delta}$ is an ML estimate defined by the product model.

The estimated lung cancer deaths per 100 000 man-years in Row $j$ and Column $k$ are given by Fit = Age fit × Smoking effect = $\exp(\hat{\mu} + \hat{\alpha}_j + \hat{\delta}_k)$.

and $t$ = years of smoking/42.5. Doll (1971) suggested that the hazard rate is approximately proportional to $d$ and to the fourth power of the duration of smoking (i.e. $\theta = 1$ and $\beta = 4$). This is a Weibull hazard function with one parameter $\beta$ independent of $d$ and $t$ and the other parameter a function of dose rate: $\gamma + \alpha d^{\theta}$, where $\gamma$ represents the background (nonsmoker) incidence at age 62.5 and $\alpha d^{\theta}$ describes the effect of smoking on lung cancer death rates. The model (3.2) with $\gamma = 0$ is equivalent to Equation (2.3) of Whittemore and Altshuler (1976) and can easily be expressed in a log-linear form,

$$\lambda(\mathbf{X}_i, \boldsymbol{\beta}) = \exp(\mathbf{X}_i\boldsymbol{\beta}) = \exp(\log \alpha + \beta \log t_i + \theta \log d_i), \tag{3.3}$$

where $\mathbf{X}_i = (1, \log t_i, \log d_i)$. Columns 3 and 2, respectively, of Table 3, give ML estimates, and their standard deviations and eye estimates, obtained by Whittemore and Altshuler for this model. When the nonsmoker data (i.e. $d = 0$) in Column 2 are included in the analysis, we require $\gamma > 0$, and for estimation purposes we use

$$\lambda(\mathbf{X}_i, \boldsymbol{\beta}) = \{\exp(\beta_2 + \beta_3 x_{i2}) + \exp(\beta_4)\}\exp(\beta_1 x_{i1}), \tag{3.4}$$

where $\mathbf{X}_i = (\log t_i, \log d_i)$ and $\boldsymbol{\beta}' = (\beta, \log \alpha, \theta, \log \gamma)$. This model is nonlinear in the parameters, and consequently the computational procedures designed for generalized linear models cannot be used. However, with the IRLS procedure, ML estimates can be obtained by using any program that provides the ability to solve a weighted least squares problem with weights that change on each iteration (see §4).

The ML estimates of the parameters for Model (3.4), using all of the data in Table 1, are given in the last column of Table 3. The diagonal terms from the **H** matrix (2.6) are easily computed by using $h_i = w_i P_i \mathbf{C} P_i'$, where $\mathbf{C} = (\mathbf{P'WP})^{-1}$ denotes the asymptotic variance–covariance matrix of $\hat{\boldsymbol{\beta}}$, and the elements of **P** and **W** are evaluated at the ML estimate $\hat{\boldsymbol{\beta}}$, and are given in Table 2a. Inspection of the $h_i$ values in Table 2a indicates that the data in the first and last columns are relatively more important with respect to Model (3.4). This emphasizes the importance of (i) the assignment of individuals to the correct dose-rate group

**Table 2**
*Regression diagnostics for the data in Table* 1, *obtained with the nonlinear model* (3.2)

| Years of smoking (midpoint) | Cigarettes per day | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0.00 | 5.20 | 11.20 | 15.90 | 20.40 | 27.40 | 40.80 |
| | (a) Diagonal terms from the **H** matrix ($p/n = 0.0635$) | | | | | | |
| 17.5 | .016 | .006 | .012 | .019 | .032 | .024 | .009 |
| 22.5 | .034 | .012 | .021 | .033 | .061 | .054 | .030 |
| 27.5 | .057 | .018 | .025 | .035 | .069 | .076 | .057 |
| 32.5 | .087 | .026 | .032 | .035 | .066 | .085 | .085 |
| 37.5 | .127 | .036 | .038 | .032 | .057 | .078 | .104 |
| 42.5 | .138 | .047 | .045 | .033 | .051 | .077 | .118 |
| 47.5 | .146 | .055 | .055 | .038 | .065 | .089 | .128 |
| 52.5 | .181 | .070 | .068 | .042 | .079 | .093 | .111 |
| 57.5 | .203 | .096 | .079 | .050 | .083 | .086 | .084 |
| | (b) Freeman–Tukey residuals | | | | | | |
| 17.5 | 1.34 | −0.08 | −0.19 | −0.33 | −0.54 | −0.43 | −0.17 |
| 22.5 | −00.17 | −0.22 | 0.93 | −0.82 | 0.02 | 0.11 | −.074 |
| 27.5 | −0.28 | −0.39 | 0.60 | −1.30 | −0.89 | 0.75 | −1.71 |
| 32.5 | −0.42 | −0.65 | 0.87 | 1.36 | 0.75 | 1.23 | 0.34 |
| 37.5 | −0.59 | 0.51 | −1.74 | −2.34 | −0.54 | 0.03 | 0.23 |
| 42.5 | −0.63 | 0.98 | −0.69 | −0.53 | 1.46 | 0.08 | 1.14 |
| 47.5 | −0.66 | −1.30 | −0.19 | −0.61 | 0.37 | −0.07 | 0.17 |
| 52.5 | −0.78 | 1.26 | 0.78 | −0.36 | 0.11 | −0.96 | −0.29 |
| 57.5 | 1.30 | −1.75 | 0.30 | 1.35 | 0.87 | −0.83 | −1.18 |

**Table 3**
*Estimates of parameters for data in Table* 1

| Parameter | Smokers | | | Smokers + nonsmokers |
|---|---|---|---|---|
| | Whittemore–Altshuler | ML | | ML |
| $\beta$ | 4.68 | 4.50 | 4.50 | 4.46 |
| | | (0.34) | (0.34) | (0.33) |
| $\log \alpha$ | 2.46 | 2.20 | 2.15 | 1.82 |
| | | (0.53) | (1.45) | (0.66) |
| $\theta$ | 1.10 | 1.18 | 1.20 | 1.29 |
| | | (0.17) | (0.40) | (0.20) |
| $\log \gamma$ | — | — | 0.96 | 2.94 |
| | | | (25.4) | (0.58) |

and (ii) the value of $d_k$ that is associated with each group (especially the last group). The FT residuals in Table 2b are used to identify outlying observations, and in this example only one FT residual exceeds 2 in absolute value. Freeman-Tukey residuals are recommended in this situation since the estimated number of deaths is small in a number of cells (24 of the $\hat{\mu}_i$ values are less than 1). The results in Table 2 do not indicate any problems for this example, but do emphasize the importance of the data on nonsmokers. These results are presented to demonstrate that the basic building blocks (see §2) required for regression diagnostics are readily available in standard statistical packages that support the IRLS procedure.

The deviance

$$D(\hat{\boldsymbol{\beta}}) = 2\{L(\hat{\boldsymbol{\beta}}) - L(\mathbf{y})\},$$

(see Nelder and Wedderburn, 1972), where $L(\mathbf{y})$ denotes the value of the log likelihood function (2.4) evaluated at $\mu_i = y_i$, $i = 1, \ldots, n$, provides an absolute measure of residual variation and is asymptotically distributed as a chi square with $n - p$ degrees of freedom (df). The deviance for Model (3.4) is 59.58 with 59 df (see the Poisson ANOVA in Table 4). The difference of the deviance between Rows 3 and 4 is used to obtain a likelihood ratio test of the hypothesis that lung cancer risk is proportional to dose rate, i.e. $H_0$: $\theta = 1$, and we conclude that this hypothesis cannot be rejected at the .05 significance level. The ML estimate of $\gamma$, the background lung cancer death rate at age 62.5, is $\hat{\gamma} = \exp(\hat{\beta}_4) = 18.9$ cases per $10^5$ man-years, and an approximate 95% interval estimate is (6.1, 59). The hypothesis $\gamma = 0$ is clearly not acceptable since the deviance goes to $+\infty$ [see (2.4)] as $\lambda$ goes to 0 for the two age groups in Column 1 of Table 1, where $y_i > 0$ and $d_i = 0$. The results of fitting Model (3.4) when the data on nonsmokers are excluded from the analysis are given in Column 4 of Table 3, and the deviance for this model is 48.2666 with 50 df. The deviance obtained by

**Table 4**
*Poisson ANOVA for data in Table* 1

| Model | Number of parameters | Residual variation | df |
|---|---|---|---|
| $\gamma$ | 1 | 445.10 | 62 |
| $\gamma t^{\beta}$ | 2 | 180.82 | 61 |
| $(\gamma + \alpha d)t^{\beta}$ | 3 | 61.84 | 60 |
| $(\gamma + \alpha d^{\theta})t^{\beta}$ | 4 | 59.58 | 59 |
| $\exp(\mu + \alpha_j + \delta_k)$ | 15 | 51.47 | 48 |

The deviance $D(\hat{\boldsymbol{\beta}}) = 2\sum_i \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$ is used as a measure of residual variation.

using (3.3) is 48.2682 with 51 df, which indicates that the background parameter cannot be reliably estimated without the control group (note that the estimated standard deviation for $\log \gamma$ is 25.4). The ML estimate of the background rate is 2.6 deaths per $10^5$ man-years, which is seven times lower than the estimate obtained when the data for nonsmokers are included. Whittemore and Altshuler (1976, p. 806) stated that nonsmokers experience a background smoking rate of 0.3 cigarettes per day, which also leads to age-specific estimates that are lower by a factor of six than those obtained when the data for nonsmokers are included.

## 4. Discussion

In this paper, we have shown that the analysis of failure rates can be viewed as a regression problem and that well-known procedures for describing, fitting and evaluating goodness of fit of models (both linear and nonlinear) can be used in the analysis of this type of data. When the dependent variable follows the Poisson distribution, the IRLS algorithm yields ML estimates of the parameters, provided that the weights are defined as described in §2 and that the convergence criteria used are equivalent to those used in the method of scoring. The deviance can be used as a measure of residual variation to develop an ANOVA-like table (see Table 4) for Poisson distributed data, but considerable care is needed when interpreting the deviance (or other measure of residual variation). First, it should be noted that the reduction in the deviance due to adding one parameter (or a group of parameters) is order dependent. Second, when the fitted values in a number of cells are small, the appropriateness of using $n - p$ as the number of degrees of freedom for a chi square test of goodness of fit is in doubt. Further, it should be noted that in observational studies with three or more explanatory variables the occurrence of cells with $c_i = 0$ (as well as cells with $c_i > 0$ but $y_i = 0$) will create problems when attempting to fit factorial-type generalized linear models with interaction terms. One should proceed with caution both when fitting these models and when constructing and interpreting the resulting ANOVA table.

Log-linear models can be easily fitted to Poisson distributed data by using standard options in statistical packages such as GLIM. For models that are nonlinear in the parameters the IRLS procedure can be used to obtain ML estimates by using any program that provides the ability to solve a weighted least squares problem with weights that change on each iteration. This can be done with GLIM, for example, by disregarding the standard options and writing macros that implement the IRLS procedure described in §2. Identical results can be obtained by using the computer program PREG (Frome, 1981) which requires a user-supplied FORTRAN function for nonlinear models. In both cases the additional input required consists of the model-specific partial derivatives $P_i$ [see (2.2)] which play the role of the predicator variables on each iteration, and initial estimates of the parameters. Listings of the FORTRAN function and the GLIM program that were used in this analysis are available.

One important area of application of Poisson regression models is the analysis of survival-time data. Poisson rate analysis can be used to analyze censored survival time data when the survival time is discrete and the explanatory variables are categorical or obtained by grouping continuous variables (which can be time-dependent). This approach is especially useful in long-term follow-up studies with large numbers of individuals, as was illustrated in §3. In this type of study, variable entry time, lost to view, and censoring are handled by using 'person-years'. An alternative approach to the problem is to assume that $c_{jk}$ is the number of individuals at risk (rather than person-years) at the beginning of the $j$th time interval, and that $y_{jk}$ follows the binomial distribution (see Prentice and Gloeckler, 1978; Pierce, Stewart and Kopecky, 1979). This leads to a partial binomial likelihood function (see Cox, 1975) with a failure probability $\lambda(\mathbf{X}_i, \boldsymbol{\beta})$, and the IRLS can be used to obtain the ML estimates of $\boldsymbol{\beta}$ by using binomial weights on each iteration. Thompson (1981) has noted that models of this type can easily be fitted with GLIM when the failure probabilities can be represented with a

complementary log–log link function, i.e. $\lambda(\mathbf{X}_i, \boldsymbol{\beta}) = 1 - \exp\{-\exp\mathbf{X}_i \boldsymbol{\beta})\}$. The IRLS procedure can be used to extend this capability to an arbitrary (nonlinear) regression model.

## Acknowledgements

The author would like to thank Dr C. C. Lushbaugh for his continuing support in this effort. Comments by the Associate Editor and the referees on an earlier draft were appreciated and taken into account in the revision of this paper. All of the computations were done using the DOE computer resources operated by Computer Sciences at Oak Ridge National Laboratory. This research was supported in part by Contract No. DE-ACO5-76R00033 between the U.S. Department of Energy, Office of Health and Environmental Research, and Oak Ridge Associated Universities, and by contract W-7405-eng-26 between the U.S. Department of Energy and Union Carbide Corporation.

## Résumé

On considère des modèles dans lesquels la vitesse à laquelle apparaissent les évènements peut être représentée à l'aide d'une fonction de régression qui décrit la liaison entre les variables prédictives et les paramètres inconnus. Des estimations des paramètres peuvent être obtenues par une méthode des moindres carrés pondérés itérative (IRLS). Quand les évènements qui intéressent suivent une distribution de Poisson, l'utilisation de l'algorithme IRLS est équivalente à celle de la méthode des scores pour obtenir les estimations du maximum de vraisemblance (ML). Le modèle de régression de Poisson général inclut des modèles log linéaires, quasi-linéaire et des modèles intrinsèquement non linéaires. L'approche présentée permet de se concentrer sur la description de la liaison entre la variable dépendante et les variables indépendantes à travers le modèle de régression. Les logiciels statistiques standards qui permettent IRLS peuvent être utilisés pour obtenir les estimations du maximum de vraisemblance, leur matrice de variances–covariances asymptotiques et les aides à l'interprétation qui peuvent être utilisées pour détecter les réponses aberrantes et les points extrêmes dans l'espace du modèle. Des applications de ces méthodes à des études épidémiologiques, les données étant organisées en forme de table de survie, sont discutées. La méthode est illustrée avec un modèle non linéaire, dérivé de la théorie de la génèse du cancer en plusieurs étapes, pour analyser la mortalité par cancer du poumon parmi les médecins anglais fumant régulièrement la cigarette.

## References

Aitkin, M. and Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics* **29,** 156–163.

Armitage, P. (1966). The chi-square test for heterogeneity of proportions after adjustment for stratification. *Journal of the Royal Statistical Society, Series B* **28,** 150–163.

Baker, R. J. and Nelder, J. A. (1978). *Generalized Linear Interactive Modelling (GLIM), Release* 3. Oxford: Numerical Algorithms Group.

Breslow, N. E. and Day, N. E. (1975). Indirect standardization and multivariate models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *Journal of Chronic Diseases* **28,** 289–303.

Carlborg, F. W. (1981). 2-Acetylaminofluorene and the Weibull model. *Food and Cosmetics Toxicology* **19,** 367–371.

Charnes, A., Frome, E. L. and Yu, P. L. (1976). The equivalence of generalized least squares and maximum likelihood estimation in the exponential family. *Journal of the American Statistical Association* **71,** 169–172.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62,** 269–276.

Crump, K. S., Hoel, D. G., Langley, C. H. and Peto, R. (1976). Fundamental carcinogenic processes and their implications for low dose risk assessment. *Cancer Research* **36,** 2973–2979.

Doll, R. (1971). The age distribution of cancer: implications for models of carcinogenesis. *Journal of the Royal Statistical Society, Series A*, **134,** 133–166.

Doll, R. and Hill, A. B. (1966). Mortality of British doctors in relation to smoking: Observations on coronary thrombosis. In *Epidemiological Study of Cancer and Other Chronic Diseases*, W. Haenszel (ed.). National Cancer Institute Monograph 19, 205–268. Washington: U.S. Government Printing Office.

Freeman, D. H., Jr and Holford, T. R. (1980). Summary rates. *Biometrics* **36,** 195–205.
Freeman, M. F. and Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics* **21,** 607–611.
Frome, E. L. (1981). Poisson regression analysis. *American Statistician* **35,** 262–263.
Frome, E. L. and Beauchamp, J. J. (1968). Maximum likelihood estimation of survival curve parameters. *Biometrics* **24,** 595–605.
Frome, E. L. and DuFrain, R. R. (1982). Analysis of cytogenetic dose–response data using a model derived from the theory of dual radiation action. Abstract. *Biometrics* **38,** 1117.
Frome, E. L., Kutner, M. H. and Beauchamp, J. J. (1973). Regression analysis of Poisson-distributed data. *Journal of the American Statistical Association* **68,** 935–940.
Gail, M. (1978). The analysis of heterogeneity for indirect standardized mortality ratios. *Journal of the Royal Statistical Society, Series A* **141,** 224–234.
Haberman, S. J. (1974). *The Analysis of Frequency Data.* Chicago: University of Chicago Press.
Hasselblad, V. (1981). Modeling dose response relationships for health effects data. In *Environmetrics* 81, Selected Papers, 179–194. Philadelphia: SIAM.
Holford, T. R. (1980). The analysis of rates and survivorship using log-linear models. *Biometrics* **36,** 299–305.
Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* **76,** 231–240.
Mantel, N. and Stark, C. R. (1968). Computation of indirect adjusted rates in the presence of confounding. *Biometrics* **24,** 997–1005.
Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135,** 370–384.
Osborn, J. (1975). A multiplicative model for the analysis of vital statistical rates. *Applied Statistics* **24,** 75–84.
Peto, R. (1972). Contribution to the discussion of paper by Cox. *Journal of the Royal Statistical Society, Series B* **34,** 205–207.
Pierce, D. A., Stewart, W. H. and Kopecky, K. J. (1979). Distribution-free regression analysis of grouped survival data. *Biometrics* **35,** 785–793.
Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics* **9,** 705–724.
Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34,** 57–68.
Taulbee, J. D. (1979). A general model for the hazard rate with covariables. *Biometrics* **25,** 439–450.
Thompson, R. (1981). Survival data and GLIM. Letter to the Editor. *Applied Statistics* **30,** 310.
Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM. *Applied Statistics* **29,** 268–275.
Whittemore, A. and Altshuler, B. (1976). Lung cancer incidence in cigarette smokers: Further analysis of Doll and Hill's data British physicians. *Biometrics* **32,** 805–816.
Whittemore, A. and Keller, J. B. (1978). Quantitative theories of carcinogenesis. *SIAM Review* **20,** 1–30.

*Received April* 1981; *revised April and July* 1982