ORISE-09-OEWH-0176

POISSON REGRESSION ANALYSIS OF ILLNESS AND INJURY SURVEILLANCE DATA

E. L. Frome
J. P. Watkins
E. D. Ellis

Center for Epidemiologic Research
Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA

C. H. Strader

U. S. Department of Energy

Date Published: December 2012

Prepared by
Oak Ridge Institute for Science and Education
P.O. Box 117
Oak Ridge, TN 37831-0117
managed by
Oak Ridge Associated Universities
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-06OR23100

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

The Oak Ridge Institute for Science and Education (ORISE) is a U.S. Department of Energy institute focusing on scientific initiatives to research health risks from occupational hazards, assess environmental cleanup, respond to radiation medical emergencies, support national security and emergency preparedness, and educate the next generation of scientists. ORISE is managed by Oak Ridge Associated Universities.

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The Department of Energy (DOE) uses illness and injury surveillance to monitor morbidity and assess the overall health of the work force. Data collected from each participating site include health events and a roster file with demographic information. The source data files are maintained in a relational data base, and are used to obtain stratified tables of health event counts and person time at risk that serve as the starting point for Poisson regression analysis. The explanatory variables that define these tables are age, gender, occupational group, and time. Typical response variables of interest are the number of absences due to illness or injury, i.e., the response variable is a count. Poisson regression methods are used to describe the effect of the explanatory variables on the health event rates using a log-linear main effects model. Results of fitting the main effects model are summarized in a tabular and graphical form and interpretation of model parameters is provided. An analysis of deviance table is used to evaluate the importance of each of the explanatory variables on the event rate of interest and to determine if interaction terms should be considered in the analysis. Although Poisson regression methods are widely used in the analysis of count data, there are situations in which over-dispersion occurs. This could be due to lack-of-fit of the regression model, extra-Poisson variation, or both. A score test statistic and regression diagnostics are used to identify over-dispersion. A quasi-likelihood method of moments procedure is used to evaluate and adjust for extra-Poisson variation when necessary. Two examples are presented using respiratory disease absence rates at two DOE sites to illustrate the methods and interpretation of the results. In the first example the Poisson main effects model is adequate. In the second example the score test indicates considerable over-dispersion and a more detailed analysis attributes the over-dispersion to extra-Poisson variation. The R open source software environment for statistical computing and graphics is used for analysis. Additional details about R and the data that were used in this report are provided in an Appendix. Information on how to obtain R and utility functions that can be used to duplicate results in this report are provided.

Key words: Poisson regression, surveillance, health event, main effects model, over-dispersion

INTRODUCTION

The mission of the Department of Energy's (DOE) Illness and Injury Surveillance Program (IISP) is to monitor morbidity and assess the overall health of the work force and to identify groups that may be at increased risk for occupation-related injury or illness. The program provides a focus for interventions that reduce or eliminate risk. The IISP also provides a means by which the effectiveness of corrective actions can be measured. This is accomplished through the routine collection, analysis, and interpretation of selected morbidity, demographic and occupational exposure data on an annual basis for each of the sites that participate in the program. To address issues of privacy and confidentiality, no identified worker data are ever transmitted off site. All data transmitted to the Program's data center are accompanied only by encrypted identifiers, and only site personnel who are directly involved with the IISP at each participating site can identify data for an individual at their site using these identifiers. In 2006 the IISP assessed the overall health of about 79,000 contractor workers at 13 DOE sites in the U.S. DOE IISP Worker Health Summary, 1995-2004 — see Strader and Richter (2007). The Worker Health Summary and annual surveillance reports for the participating sites are available at DOE's Office of Health, Safety, and Security internet site, or from the Office of Scientific and Technical Information (OSTI).

This report reviews regression methods for the analysis of data when the response variable is a count, and describes how these methods can be used for the analysis of the IISP data. The general count regression approach can be applied to any situation in which the response variable is non-negative integer valued with expectation that varies as a function of known covariates. The IISP data are collected by coordinators at each site and submitted to the Epidemiologic Surveillance Data Center located at the Oak Ridge Institute for Science and Education (ORISE), where quality control and record linkage procedures are carried out. The source data files obtained from each facility are i) a roster, ii) a return to work (RTW) absence file, and iii) an OSHA-Recordable events file. Workers absent five or more consecutive workdays due to any illness or injury are cleared to return to work through the occupational medical clinics at each site (DOE 10 CFR851). The latter two files contain "health event" records. There is at least one health event, i.e., a disease or injury diagnosis, associated with each health event record. The event files and the roster are maintained in a relational data base and are used to generate stratified tables that serve as the starting point for Poisson regression analysis. The explanatory variables that are used to define these tables are age, gender, occupational group and time. Age at risk is represented as a factor with four levels (<30, 30-39, 40-49, 50 or greater) and there are seven occupational groups. Time is in one year intervals and the number of levels depends on the length of time the site has been in the program. This choice of explanatory variables is based on programmatic and practical considerations. Typical health events are respiratory disease, circulatory disease, and injuries, but any disease of interest as defined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) (National Center for Health Statistics and Centers for Medicare & Medicaid Services, 2011) code can be specified as the response variable. Each of the cells in the stratified table contains the number of events and person-time at risk for workers with the corresponding characteristics. For the RTW data, an event is an absence of at least five consecutive workdays with an illness or injury as defined by the ICD-9-CM coding system, e.g., for respiratory illness the ICD-9-CM code is 460-519. The event rate is the number of events divided by the person-time at risk in each stratum and is expressed in units of 1000 person-years, i.e., if y is the number of events and n is the person-time at risk (in units of 1000), then r = y/n is the event rate for the stratum.

The number of events in a given stratum is the dependent, or response, variable and is assumed to follow a Poisson distribution. The mean of the distribution is equal to the person-time at risk multiplied by the "rate function" which depends on the level of each of the explanatory variables. In the IISP, occupational group and year are viewed as "exposure" variables of primary interest and age and gender are of secondary interest as potential confounding variables and/or effect modifiers. The underlying rate function is represented by a "product model" in which the stratum specific rates are the product of parameters, one for each level of the explanatory variables (subject to some constraints). For purposes of estimation the product model is re-parameterized as a log-linear model and the log scale parameters are estimated using standard options in the generalized linear model (GLM) function in the open source program R (2009). In the log-linear parameterization the resulting model is equivalent to a "main effects model" (MEM), and Poisson regression is used to obtain maximum likelihood (ML) estimates of the parameters, their standard errors, and other relevant information. The results are presented in a summary table and a "main effects plot". The summary table for the MEM includes parameter estimates, standard errors, estimates of the dispersion parameter, and a score tests for "over-dispersion". The summary table and the main effects plot facilitate the evaluation of the event rates for each event type of interest from the RTW and OSHA data bases for each of the thirteen sites. An analysis of deviance table is obtained to evaluate the importance of each of the explanatory variables and each of the possible two factor interactions. The two factor interactions may be of interest if there is an indication of over-dispersion for the MEM since over-dispersion may be due to lack of fit of the MEM and/or heterogeneity of variance of the counts under the Poisson assumption.

The first application of Poisson regression was given by Cochran (1940) with wireworm counts from an agriculture experiment as the response variable and the regression function $E(y_i) = \mu_i = (x_i \beta)^2$, where x_i is the i^{th} row of the model matrix for a Latin square design. This model is a GLM with a Poisson response and a "square root" link function — see Frome (1984) for additional details. Jorgenson (1961) proposed Poisson regression with a linear rate function for use in consumer demand analyses and reliability. Nelder and Wedderburn (1972) described GLM for response variables in the regular exponential family — see also McCullagh and Nelder (1989, Chapter 6) for Poisson log-linear models. Frome et al. (1973) described Poisson regression methods for general models with an emphasis on intrinsically nonlinear models. Frome (1983) described the analysis of event rates for Poisson data and Frome and Checkoway (1985) considered applications of these methods in epidemiologic follow-up studies. Breslow and Day (1987) described the use of Poisson regression in occupational cohort studies and Koch et al. (1986) gave a more complete review of Poisson regression methods and areas of applications. Cameron and Trivedi (1986) discussed Poisson regression in econometric applications. Richardson and Loomis (2004) reviewed the use of Poisson regression in occupational and environmental cohort studies and considered problems that may occur when person-time and events are tabulated by levels of an exposure variable that was originally measured on a continuous scale and has been categorized for analysis.

Wing et al. (1991) and Frome et al. (1997) have applied these methods to mortality studies of nuclear industry employees. Poisson regression has also been used extensively in mortality studies of atomic bomb survivors — see e.g. Pierce et al. (1996). Preston et al. (1993) have developed special purpose software that supports the use of Poisson regression in the atomic bomb survivors' studies and other situations that require Poisson regression with excess relative risk models. Poisson regression has been used extensively in the analysis of motor vehicle accidents — see e.g. Erlander et al. (1972), Frome and Walton (1975), Gustavsson and Svensson (1976), Michner and Tighe (1992), Fridstrom et al. (1995), Li (2001), Lord et al. (2005).

The application of Poisson regression to the analysis of occupation injuries in various situations has also been considered. Mohr and Clemmer (1989) evaluated occupational injury intervention programs in the petroleum drilling industry. Mallick and Mukherjee (1996) studied accidents in mines; Bailer et al. (1997) presented a case study of injury rates in workers in agriculture, forestry, and fishing; and Richardson et al. (2004) evaluate fatal injury rates by race and ethnicity in southern and non-southern states. Smitha et al. (2001) used Poisson regression to evaluate the effect of mandatory state workplace safety interventions on occupational injury rates during a five-year time interval. Various economic, regulatory, and demographic covariates were included in their model. Inclusion of these additional explanatory variables was important since state-to-state differences can be problematic when comparing occupational injury rates. Melchlor et al. (2005) used Poisson regression models to evaluate sickness absences from work in the French GAZEL cohort study. Absence rates in occupational classes were of primary interest and were adjusted for other explanatory variables (age, demographic characteristics, and health behaviors). Karra (2005) used Poisson regression and related regression models to evaluate fatality and injury rates among operator and contractor employees in the underground and surface mining of various minerals. Laaksonen et al. (2008) used Poisson regression to evaluate gender differences in sickness absence rates on municipal employees from Finland. Explanatory variables included in the analysis were age, occupational class, health status measures, and working conditions

In each of these situations, the dependent (response) variable is a count that may follow the Poisson distribution with a mean that varies as a function of several explanatory variables. A regression function describes the relation between the count and the explanatory variables, and for a fixed set of explanatory variables the mean of the count variable is equal to the regression function times the person time at risk. Consider, for example, the number of events that occur in a group of workers that are classified by gender, age, and workplace characteristics over a specified period of time. Suppose that x_i represents the value of these explanatory variables for the i^{th} group with n_i workers in the group and y_i events occur. The expected value of y_i is equal to $\mu_i = n_i \lambda(x_i, \beta) = n_i \lambda_i$, where $\lambda_i = \lambda(x_{i,}\beta)$ is a known function and β is an unknown vector of parameters. As a general rule the n_i are much larger than the y_i (although this is not required), and λ_i is the event rate in the i^{th} group. The regression function $\lambda(x_i, \beta)$ represents the systematic variation in event rates that can be described by the explanatory variables.

It was first observed by Bortkiewicz (1898) in his study of death by horse-kick in the Prussian army that the variation in y (with x_i constant) can be described by the Poisson distribution, i.e., $f(y|\mu) = \exp(-\mu)\mu^y/y!$ — see also Preece et al. (1988) and Quine and Seneta (1987). Haight (1967, Chapter 9) described the historical development of the Poisson and related distributions from 1781 to 1920 and notes that although Poisson discovered the mathematical expression, Bortkiewicz discovered its use as a probability distribution for discrete data. He further claimed that the Poisson distribution is second in importance to the normal from both a theoretical and applied perspective. Haigh (1967, Chapter 7) reviewed a number of areas where the Poisson distribution has been used to describe the distribution of counts based on both empirical and theoretical studies. The important point to note here is that in the evaluation of health events of interest the variation in the counts y_i (for fixed x_i) is described by the Poisson distribution. For example if there are say $n_i = 625$ workers in the i^{th} group and the event rate is $\lambda_i = 4/1000$, then y_i will follow the Poisson distribution with mean 2.5. Consequently, under the Poisson model if a large number of groups with the same characteristic were observed, the number of events would vary from zero approximately 8% of the time to as high as ten, and two events would occur most frequently. Even though the relation and factors that describe the event rate potential are known, the number of events is "random" in character, and we say that this "unexplained" variability is due to Poisson variation. Methods for testing the model assumptions concerning the regression function and Poisson variation are presented, and alternative methods are described for situations in which either or both of the assumptions is in question.

2. STATISTICAL METHODS

In the IISP a response variable of interest is the number of events that occur during a specified period of time. The number of events and corresponding time at risk are obtained for each combination of levels of the explanatory variables (also referred to as covariates, dependent variables, exogenous/endogones variables, and predictor variables). In general, covariates can be continuous (e.g. age) or categorical (e.g. gender). In large observational studies such as IISP it is often convenient to use "grouped data" methods, i.e. age is categorized into four intervals and time is in one year intervals. The use of grouped data methods in the analyses of rates and rate standardization is described by Breslow and Day (1987, Chapter 2). Rostgaard (2008) presents a detailed discussion of computational issues involved in the creation of the event-time tables required for the grouped data methods. The algorithm first described by Clayton — see Breslow and Day (1987, Appendix IV) — was used as the basis for several macros coded in SAS (2008), and the relationship to similar methods implemented in several other commercial packages and R is discussed. Clayton and Hills (1993, Chapters 5 and 6) provide a detailed discussion of rate estimation in epidemiologic studies and the use of Lexis diagrams to obtain stratified tables in situations when multiple time scales (i.e. age and calendar time) are of interest. They discuss the relationship between the instantaneous probability rate for an individual at a specified point in time and frequency of events in a group of individuals observed over time. The estimated rate parameter is based on the experience of a group of workers who are assumed to have the same rate parameter over a specified period of time. The event rate is obtained by dividing y, the number of events, by n, the (person) time at risk, for each possible combination of levels of all of the explanatory variables. The Poisson regression model is based on the assumption that the event counts follow a Poisson distribution with expected value equal to the person time at risk times an underlying rate, i.e., $E(y_{ijkt}) = n_{ijkt}\,\lambda_{ijkt}$ for each of the possible combinations of the explanatory variables. When the rate function λ is multiplicative (so that the "link" function is logarithmic) and the explanatory variables are categorical, a standard notation for a log-linear Poisson regression model with explanatory variables age, gender, occupational group, and year is:

$$\log(\lambda_{ijkt}) = \mu + \alpha_i + \gamma_j + \theta_k + \tau_t \,. \qquad (1)$$

This MEM includes the four explanatory variables and no "interaction" terms. In equation (1) the α_i represent age effects, the γ_j represent gender effects, the θ_k represent occupational group effects, and the τ_t represent year effects, all on the log scale. The MEM is over-parameterized and the interpretation of μ (often referred to as the "intercept term") depends on the procedure used to eliminate redundancy (see the Examples). For practical reasons it is sometimes more convenient to write equation (1) as

$$\log(\text{rate}) = m + ag + sx + og + yr \qquad (2)$$

where
m is the intercept term,
ag represents age group with 4 levels ag[i] , i = 1,2,3,4 ,
sx represents gender with sx[j] = 1 for females and sx[j] = 2 for males ,
og represents occupational group k = 1,...,7 (see Table 1 for names of occupational groups), and
yr represents calendar year t=1 ,...,T (T depends on the DOE site)

This notation corresponds to that used in statistical programs that fit Poisson regression models (e.g. GLIM, S-Plus, R). For example, in computer output the value of og[k] is the ML estimate of the log scale effect for the kth occupational group (see Example 1 and Appendix 2 for details).

For mathematical and notational convenience, let $r_i = y_i/n_i$ denote the event rate in the ith group and consider the following weighted sum of squares

$$S(\beta) = \Sigma_i\, w_i\, [r_i\ -\lambda(x_i,\beta)]^2\ , \qquad (3)$$

where the weight w_i is inversely proportional to the variance of r_i. For the MEM x_i is based on a row vector of “indicator variables” that is defined by the levels of the four explanatory variables in the ith cell and $\lambda(x_i,\beta) = \exp(x_i\beta)$, where $\beta = (\alpha,\gamma,\theta,\tau)'$ is a vector that contains the unknown parameters as described in equation (1). In matrix notation the log of the rate function is $\mathbf{X}\beta$ where $\mathbf{X}$ is an N by q matrix with rows x_i, N is the number of cells in the stratified table, and q is the number of estimable parameters.

If the y_i are independent and follow the Poisson distribution the weights in equation (3) are $w_i=n_i/\lambda(x_i,\beta)$. Since $\lambda(x_i,\beta)$ is, in general, nonlinear in the unknown parameters and the w_i depend on β an iterative procedure can be used to obtain an estimate of β. The rate function $\lambda(x_i,\beta)$ in (3) is replaced with the linear terms in a Taylor series expansion about an initial estimate $\beta°$ and the resulting weighted least squares system of equations is solved for a correction vector $\delta°$. The initial estimate is updated, i.e. $\beta^1 = \beta°+\delta°$ and the iterative weighted least squares (IWLS) continues until some convergence criteria are satisfied — see Frome et al, 1973, Frome (1983) for details. Under the Poisson assumption the IWLS procedure is equivalent to using the method of scoring to find the ML estimate of β where the kernel of the log-likelihood function is $L(\beta) = \Sigma_i[y_i\log\{n_i\lambda(x_i\beta)\}- n_i\lambda(x_i,\beta)]$.

Maximizing $L(\beta)$ is equivalent to minimizing the deviance

$$D(\beta) = 2[L(y) – L(\beta)] = 2\ \Sigma_i[y_i\log(y_i/\mu_i) – (y_i -\mu_i)], \qquad (4)$$

where $\mu_i = n_i\lambda(x_i,\beta)$ and $L(y)$ denotes the value of the log likelihood evaluated at $\lambda_i = y_i/n_i$. This is called the saturated model since there is a parameter for each cell in the table. Note that the IWLS procedure does not require calculation of the log-likelihood function since convergence can be defined in terms of the relative change of the parameter estimates (Frome et al 1973). If a stable solution $\hat{\beta}$ is found, it will be the root of the likelihood equations, i.e. $L(\hat{\beta})$ will be at its maximum and $D(\hat{\beta})$ will be at its minimum value (Charnes et al, 1976). The IWLS algorithm does not minimize the weighted sum of squares in equation (3) but rather is a solution to the estimating equations and consequently can be used for quasi-likelihood and resistant alternatives to ML — see

Green (1984) for a more detailed discussion of the IWLS procedure. When $\lambda(x, \beta)$ is a generalized linear function, i.e., $\lambda(x_i, \beta) = \exp(x_i\beta)$, and the response variable y is in the exponential family, the result is a GLM. Nelder and Wedderburn (1972) showed that for GLMs the IWLS procedure is equivalent to ML. This was the basis for the statistical package GLIM (Baker and Nelder, 1978). There are currently a number of statistical packages that support the analyses of count regression models ─ see e.g. Data Analysis Examples (2012) at the Statistical Computing website at UCLA Academic Technology Services, where five statistical packages that can be used to fit Poisson and other count regression models are described and illustrated with examples. In this report the R (2012) open source language and environment for statistical computing is used for all data analysis, e.g. the R function **glm**( ) is used to fit Poisson log-linear models.

The deviance (4) provides an absolute measure of residual variation and is asymptotically distributed as a chi-square with N-q degrees of freedom (df). The difference of the deviance for nested models follows the chi-square distribution and is used to obtain an analysis of deviance table. Also included in the analysis of deviance table is the value of the information criteria (IC)

$$IC = D + kq\varphi \,, \tag{5}$$

where D is the deviance, k is a constant, q is the number of estimable parameters, and φ is a dispersion parameter. For the Poisson model $\varphi = 1$ and if $k = 2$ then equation (5) is known as the Akaike information criteria. Atkinson (1981) suggests that k should range from 2 to 6. McCullagh and Nelder (1989, Chapter 3.9) discuss the use of the IC in model selection and, following their suggestion, we use $k = 4$ unless stated otherwise. This corresponds to use of the 5% point of the t (or F) distribution for a covariate with one df. The IC is an optimality criterion that is used (with smaller values being preferred) to compare models. Adding a term to a model will always decrease the deviance, so adding k times the number of parameters is a penalty for unnecessary terms. The IC can be used in a model selection procedure that implicitly fits all possible hierarchical models and identifies good models for further consideration (Ostrouchov and Frome, 1993).

The deviance can also be used to calculate a pseudo R-squared measure for Poisson models, $R^2_D = 1 - D(\hat{\beta})/D(\hat{\lambda}_o)$, where $D(\hat{\lambda}_o)$ denotes the deviance for the minimal one parameter model $E(y_i) = n_i(\lambda_o)$. R^2_D represents the relative reduction in the deviance due to the covariates in the model and is similar to R^2 measures used in linear regression. If the sample size is small relative to the number of parameters in the full model R^2_D may be inflated. A bias adjusted R-square measure for Poisson and quasi-poisson regression is given by $R^2_A = 1 - [D(\hat{\beta}) + q\text{-}1]/D(\hat{\lambda}_o) = AR^2_D$, where A is the "shrinkage factor" — see Heinzland and Mittlbock (2003) for further details.

After reviewing the results from fitting a Poisson regression model, some type of model checking may be indicated. There are a number of model-checking techniques that can be considered if there is some indication that the Poisson model is inadequate. These techniques can be both formal and informal and usually involve the analysis of residuals, regression diagnostics, and the fitting of more complex models. Model checking methods are described in detail for count data by Cameron and Trivedi (1998, Chapter 5) and for GLMs by McCullagh and Nelder (1989, Chapter 12). The two primary ways that the results obtained from Poisson regression may require further evaluations are lack-of-fit of the regression function (in this situation, the MEM) and extra-Poisson variation. In either or both situations the counts may display over-dispersion relative to the Poisson model. Pregibon (1981) first proposed that when binomial logistic regression methods are used in observational studies diagnostic procedures similar to those used in standard linear regression should be used to check for outlying y values (as indicated by large values of the standardized residual) and extreme points in the model space. Correspondingly, an informal procedure that is used to check for systematic departures from the Poisson MEM is based on plots of residuals versus the fitted values from the MEM. Following McCullagh and Nelder (1989, Chapter 12) we plot the standardized deviance residuals (SDR) against the fitted rates from the MEM (see Section 3.2 for an example). The SDR for the i^{th} observation is obtained by dividing the signed deviance residual sign(y_i - μ_i)d_i by $\sqrt{1-h_i}$, where $\hat{\mu}_i = n_i \hat{\lambda}_i = n_i \lambda(x_i, \hat{\beta})$, $d_i = 2[y_i \log (y_i/\hat{\mu}_i) – (y_i - \hat{\mu}_i)]^{1/2}$, and h_i is the i^{th} diagonal element from the leverage or “hat matrix” H — see Frome (1983). Note that d_i^2 is the value of the deviance in equation (4) for the i^{th} observation evaluated at the ML estimate $\hat{\beta}$. In this plot the area of each square is proportional to the final weight $w_i=n_i/\lambda(x_i, \hat{\beta})$ in equation (3). The SDRs outside the 99% limits are marked with a red x. A second useful plot for model checking purposes is a normal Q-Q plot of the SDRs. This plot is used to identify extreme values which would appear in the upper right and/or lower left portion of the plot. McCallaugh and Nelder (1989) point out that residuals from data with many zero counts could result in many small residuals near zero which may appear as a plateau in the Gaussian Q-Q plot. A third diagnostic plot is obtained by plotting the SDRs versus the scaled $\tilde{h}_i$ values $\tilde{h}_i = nh_i/q$. Thus $\Sigma \tilde{h}_i = n$ since $\Sigma_i h_i = q$, and $\tilde{h}_i > 2$ indicates high leverage (McCallagh and Nelder, 1989, Section 12.7). The outlying points are identified in this residual-leverage plot by a red “x” at points with both $\tilde{h}_i > 2$ (i.e. to the right of the solid vertical line) and SDRs that are outside the horizontal dashed lines at +/- 2.58 (the 99% limits). A fourth diagnostic plot of the absolute value of the SDRs against the fitted values gives an informal check of the adequacy of the assumed variance function. The null pattern will not show a trend, and smoothing is used to identify a possible pattern. These four plots are illustrated in Section 3.2, and these and additional diagnostic plots for Poisson models available in R are discussed by Maindonald and Braun (2007, Section 8.4).

One of the main purposes of the diagnostic plots is to identify “outliers” and/or over-dispersion. If outliers are not considered to be a problem then extra-Poisson variation is a potential reason for over-dispersion. Over-dispersion is a common complication in the analysis of count data using Poisson regression. The occurrence of excess variation has little effect on the regression coefficients of primary interest, but can result in serious errors in estimated standard errors, test statistics, and confidence intervals. One approach to dealing with over-dispersion is to consider a random effects mixed Poisson model in which the y_i follow a Poisson distribution with mean $v_i\mu_i$ where the v_i are continous independent positive-value random variables with mean one and finite variance. If the distribution of y_i given v_i and x_i is Poisson ($v_i\mu_i$), then the marginal mean and variance of y_i are

$$\mu_i = n_i \lambda(x_i,\beta) \text{ and } \mathrm{var}(y_i) = \mu_i + \delta \mu_i^b , \qquad (6)$$

where b = 1 or 2 and $\delta > 0$ is often referred to as the index of dispersion parameter. If the v_i follow the gamma distribution then the y_i follow the negative binominal (NB) distribution (Lawless, 1987) and NB regression can be used to obtain ML estimates of the regression parameters β and the dispersion parameter δ. The Poisson assumption can be tested using a likelihood ratio statistic for the null hypothesis $\delta = 0$. In this situation $\delta = 0$ is on the boundary of the parameter space and the likelihood ratio test statistic has an asymptotic distribution with a probability mass of ½ at 0 and a ½ $X^2(1)$ distribution above 0. Lawless (1987) indicates that the asymptotic distribution of the likelihood ratio and related statistics substantially overestimate the significance level associated with these tests and indicates that a score test should be considered to test for extra-Poisson variation. Dean and Lawless (1989) review various aspects of the mixed Poisson model with b = 2 in equation (6) and derive the following score test statistic for extra Poisson variation

$$T_a = \Sigma_i\{(y_i - \hat{\mu}_i) - y_i + h_i\hat{\mu}_i\} / (2\,\Sigma_i\hat{\mu}_i^2)^{½} . \qquad (7)$$

Large positive values of T_a indicate over-dispersion and for large datasets (n > 50) T_a is approximately distributed as standard normal variable. T_a requires only the Poisson model be fitted to obtain $\hat{\mu}_i$ and h_i in (7). Dean (1992) further describes T_a and two additional tests for over-dispersion both of which are designed to be powerful against arbitrary alternative mixture models where only the first two moments of the mixed distribution are specified.

An alternative to the NB regression approach to dealing with over-dispersion is to use the quasi-likelihood method of moments (QL/M) procedure (Breslow, 1990) based on the mean and variance structure as defined in equation (6). The regression coefficients are estimated using the quasi-likelihood estimating equations and the variance parameter is estimated by the method of moments. If a specific distribution for the v_i is not known, then in the most widely used QL/M procedure we assume, as an approximation, that the mean and variance function are as stated in equation (6) with b=1 so that $E(y_i)=\mu_i=n_i\lambda(x_i,\beta)$ and $var(y_i)=(1+\delta)\mu=\varphi\mu$ for some constant φ (see McCullagh and Nelder, 1989, Chapters 6 and 9). They note that even relatively substantial errors in the assumed functional form of var(y) generally have only a small effect on the conclusions. This approach is referred to here as QL/M1 regression and requires only a minor adjustment to the results from the Poisson model. For the QL/M1 regression model the estimating equations for the parameters β are the same as the ML estimating equations and are not affected by the value of φ. The moment estimator of φ is the Pearson chi-square statistic divided by its df, i.e. $\tilde{\varphi} = [\Sigma_i(y_i - \hat{\mu}_i)^2/\mu_i]/[n-q]$). The standard errors of the regression parameters are multiplied by $\sqrt{\tilde{\varphi}}$, and deviance differences are rescaled by this estimate. Approximate F tests are used in the rescaled analysis of deviance table and the t distribution is used to obtain confidence intervals for parameter estimates using the adjusted standard errors. These adjustments to account for over-dispersion using the QL/M1approach are discussed by Venables and Ripley (2002, Chapter 7) and are implemented in the R function **quasipoisson**( ).

The second QL/M procedure is based on using the quadratic variance [i.e., b=2 in equation (6)]. This corresponds to the variance of a NB regression model as described by Lawless (1987). This QL/M2 procedure was proposed by Breslow (1984) as a method for dealing with extra-Poisson variation. The QL/M2 estimates can be obtained using the IWLS procedure derived from equation (3) using weights that are based on equation (6) with b=2, i.e., $w_i = n_i / var(y_i) = n_i / (\mu_i + \delta\mu_i^2)$. This requires an initial Poisson fit with δ=0 so the weights are $w_i = n_i / \mu_i$, where μ_i is evaluated at $\hat{\beta}$. The moment equation (Breslow, 1990)

$$U(\beta, \delta) = \Sigma_i\{(y_i - \mu_i)^2 / (\mu_i + \delta\mu_i^2) - (n-q)/n\} = 0 \ , \qquad (8)$$

(with μ_i evaluated at the current estimate $\tilde{\beta}$ of β) is solved for δ. The IWLS procedure is repeated with weights $w_i = n_i / (\mu_i + \tilde{\delta}\ \mu_i^2)$, where $\tilde{\delta}$ is the current estimate of δ obtained from equation (8) and μ_i is evaluated at $\tilde{\beta}$. This procedure can be implemented using standard software. For example, using R, the Poisson regression is obtained using **glm**( ) function with the Poisson family and with prior weights

$1/(1+\tilde{\delta}\ \tilde{\mu})$, and the moment equation (8) is then solved using **uniroot**( ). If solving the estimating equations leads to a negative value of $\tilde{\delta}$, the estimate is set equal to 0 since δ is assumed to be nonnegative (Breslow, 1990). Lawless (1987) provides a detailed discussion of both NB regression and the QL/M2 procedure. Note that when δ is known, the NB distribution is in the regular exponential family, i.e. the IWLS procedure yields ML estimates of β. Consequently, if in the QL/M2 procedure, the moment equation (8) is replaced with the ML equation for δ based on the NB distribution (assuming β is known), the iterative procedure will yield the ML estimates of β and δ – see also the R function **glm.nb**( ).

3. APPLICATIONS

To illustrate the use of Poisson regression methods for the IISP data, we use respiratory events at two of the DOE sites described in the Worker Health Summary as the dependent count variable. Respiratory events include acute disorders such as the common cold, pneumonia, and influenza; chronic life-threatening diseases like emphysema and chronic beryllium disease; allergic reactions resulting from exposure to pollen, dust, and other environmental irritants. This diagnosis was the most common reason for an absence among workers in the Worker Health Summary. For each of the two sites a stratified table of respiratory events and person-years was obtained from the roster file and the RTW file for the four explanatory variables, age (4 levels) gender (2 levels), occupational group (7 levels), and years (10 levels). The resulting file for each site was in text (CSV) format and was input to R as a data frame with columns having variable names ag, sx, og, yr, rsp. Certain combinations of ag, sx, og, and yr had zero person-years (n=0) and, consequently, y (the number of respiratory events) must be zero for these cells. These "empty cells" are not included in the data frames. A data frame is the analytic data structure that is used in R to fit GLMs, and in this situation, additional health event counts would be added as columns to the data frame (see Appendix 2). There are two kinds of empty cells. Necessarily empty cells occur when a combination of factor levels is a priori impossible. Accidently empty cells occur when a combination of factors is possible but does not occur in the observed data. For cells that have n > 0 the event count is a non-negative integer and an observed zero is a "sampling zero" (see McCullagh and Nelder, 1989, Sections 3.7.1 and 6.3.2 for further details). Note that a similar situation occurs when Poisson regression is used in the log-linear analysis of sparse multi-dimensional contingency tables, and zero counts are referred to as being either "sampling" or "structural" (see e.g. Brown and Fuchs, 1983, Haslett, 1990). The terms structural zero and sampling zero are also used in the context of zero- altered Poisson regression models with a different interpretation (see Appendix 1).

3.1 EXAMPLE 1

For Site 1 the text file that was input to R had 537 rows. Of the 560 possible cells 23 had zero person-years (n=0). Table 1 shows the marginal table of respiratory events, person-years, and rates for occupation group and age at a DOE site. For Site 1 the overall (crude) five-day absence rate is 1226/73.681 = 16.64 events per 1,000 person-years.

Table 1. Respiratory Absences by Occupational Group and Age Group at Site 1

OCCUPATION	AGE			
	16-29	30-39	40-40	50+
a) Number of Absences				
A-Administrative	11	72	156	171
C-Crafts	2	17	34	31
E-Fire and Security	1	7	11	3
O-Line Operators	1	0	5	4
P-Professionals(F I M)	7	78	184	163
S-Service	2	20	35	30
T-Technical Support-B	5	37	78	61
b) Person Years				
A-Administrative	1,697	3,801	6,446	5,907
C-Crafts	95	467	1,215	1,061
E-Fire and Security	96	466	445	201
O-Line Operators	66	315	398	186
P-Professionals(F I M)	1,980	9,250	14,455	13,030
S-Service	210	342	771	758
T-Technical Support-B	583	2,330	4,124	2,986
c) Respiratory Absence Event Rate Per 1000 Person Years				
A-Administrative	6.5	18.9	24.2	28.9
C-Crafts	21.1	36.4	28.0	29.2
E-Fire and Security	10.4	15.0	24.7	14.9
O-Line Operators	15.2	0.0	12.6	21.5
P-Professionals(F I M)	3.5	8.4	12.7	12.5
S-Service	9.5	58.5	45.4	39.6
T-Technical Support-B	8.6	15.9	18.9	20.4

The results of fitting the MEM are summarized in Table 2. The first line in Table 2 shows the deviance for the minimal (one parameter) model which is 937.16 with 536 df, and the crude rate on a linear and log scale is shown on line two. The deviance, df, number of parameters (q), information criterion (IC), and the adjusted R-squared for the MEM are on line four of Table 2. The difference in the deviance for the MEM and the minimal model is 451.7 with 19 df. This is a test statistic for the hypothesis that all of the parameters for the explanatory variables in equation (1) are zero, and is compared to the chi-square distribution with 19 df (the 99.99 percentile of chi-square with 19 df is 50.8). Estimates of the Pearson chi-square statistic and corresponding dispersion parameter phi.p are on line five. The score statistic T_a (see equation 6) is also shown on line five. Under the Poisson MEM model, the dispersion parameter is assumed to be one and estimated values larger than one are taken as an indication of over-dispersion. If the Poisson MEM provides a good fit, the dispersion statistic should be close to one. Hilbe (2008, Chapter 4) provides a detailed discussion of over-dispersion and various methods for dealing with it. One formal approach to evaluating over-dispersion is to compare the score statistic T_a to the standard normal distribution (see Methods). It is important to note that over-dispersion may be due to

Table 2. Poisson Regression Results for Respiratory Absences at Site 1

Minimal Model: Deviance = 937.16, n = 537
Rate(unadjusted)= 16.64 resp events/1000, m = 2.812

Main Effects Model: log(rate)= m + ag + sx + og + yr

Deviance = 485.434, df = 517, q = 20, IC = 565.43, R^2 = 46.2%
Dispersion: phi.p = 1.37, Pearson X^2 = 707.9, Score Ta = 0.33
Rate(adjusted)= exp(m)= 17.28 resp events/1000, m = 2.85

a) Age Groups: ag[] Estimates

	<30	**30-**	**40-**	**50+**
Est_L%	-124.67	-30.48	0.00	7.35
SE_L%	19.17	8.01	0.00	6.46
RR	0.29	0.74	1.00	1.08
AdjA	4.97	12.74	17.28	18.60

b) Gender: sx[] Estimates

	F	**M**
Est_L%	72.34	0.00
SE_L%	6.45	0.00
RR	2.06	1.00
AdjG	35.63	17.28

c) Occupational Groups: og[] Estimates

	A	**C**	**E**	**O**	**P**	**S**	**T**
Est_L%	-14.77	50.74	13.90	-53.52	-60.47	70.04	-5.92
SE_L%	7.91	11.15	19.09	27.42	7.30	10.93	8.72
RR	0.86	1.66	1.15	0.59	0.55	2.01	0.94
AdjO	14.91	28.71	19.86	10.12	9.44	34.82	16.29

d) Year: yr[] Estimates

	1995	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
Est_L%	40.79	31.97	-3.36	-20.67	17.16	-13.63	-10.80	-11.03	-9.40	-21.02
SE_L%	7.24	7.57	9.14	9.59	8.16	9.33	9.21	9.06	8.88	9.22
RR	1.50	1.38	0.97	0.81	1.19	0.87	0.90	0.90	0.91	0.81
AdjY	25.99	23.79	16.71	14.06	20.52	15.08	15.51	15.48	15.73	14.01

**The Reference Group for Adjusted Rates is Males Age 40-49*

“heterogeneity of variance”, “lack-of-fit” of the regression function (i.e. misspecification of the λ_i), or both — see Frome et al, (1973), Dean and Lawless (1989). Misspecification of the regression function could occur, for example, if important interaction terms are not included in the model.

The rest of Table 2 displays the ML estimates of parameters in the MEM. The MEM is over-parameterized and two general approaches have been used to eliminate redundancy. In the first approach, a specific category is selected as the reference level, the corresponding parameter is set equal to zero, and each of the remaining parameters represents a comparison of that level with the reference level. If the levels of an explanatory variable correspond to exposure to a hazardous material with each level representing an increasing level of exposure, the first level is usually specified as the reference level — see Breslow and Day (1987, Chapter 4). This approach to eliminating redundancy, referred to as the "contr.treatment" contrast in R, is used for gender with male as the reference level. That is, in equation (1) $\gamma_2 = 0$ and γ_1 represents log of the rate ratio (RR) for female relative to male workers. This contrast is also used for the age factor with level three (ages 40-49) as the reference level, i.e. $\alpha_3 = 0$ and α_j ($j \neq 3$) is the log of the relative effect of the j^{th} age group. The age group 40-49 is representative of the workforce at IISP sites in DOE complex which has been stable but aging over time. The average age of the workers increased from 42 in 1995 to 46 in 2004 (see Worker Health Summary, Chapter 2).

The parameter μ in the MEM (see equation 1) is often referred to as the "intercept" term and represents the log of the rate at the reference level of age and gender, i.e., males age 40-49. The ML estimate m= 2.85 of μ is in Table 2 on the line above panel a. The age and gender adjusted rate, 17.28 respiratory events per 1000 person-years, is shown on the same line. The ML estimates of the ag parameters are listed in Table 2-a (on the line that begins Est_L%) in logarithmic percent (L%) units (Tornqvist et al. 1985), i.e., the ML estimate ag[4] of α_4 (workers 50 and older) is 7.35L%. The estimated standard error 6.46L% is shown below the parameter estimate on the line SE_L%. The RR for males age 50 and older compared to age 40-49 is exp(0.0735) = 1.08 and is on the next line of Table 2-a. The adjusted rate AdjA =
exp(m + ag[4]) = exp(2.85 + 0.0735) = 18.60 respiratory events per 1,000 person-years is on the last line of Table 2-a. The ML estimate sx[1] of γ_1 is 72.34L% with a standard error of 6.45L% — see Table 2(b). The corresponding RR for females is 2.06 and the adjusted rate for females age 40-49, 35.63 events per 1,000 person-years, is on the last line of Table 2-b.

A second method that is used to deal with over-parameterization in the MEM is to require that the parameter estimates sum to zero. This is called the "contr.sum" contrast in R. The sum to zero contrast is used in the analysis of multidimensional contingency tables (Bishop et al, 1975) and was used by Cochran (1940) in the analysis of counts from a Latin square design. Moore and Beckman (1988) used this constraint to evaluate the effect of five factors on the failure rate of nuclear reactor values. The sum to zero constraint is used here for occupational group, i.e. in equation (1) $\Sigma_k\theta_k = 0$ or using the equation (2) naming convention Σ_k og[k] = 0. The ML estimates og[k] of the θ_k are on the line in Table 2-c (labeled Est_L%), and the corresponding standard errors are on the line labeled SE_L%. The RRs are on the next line of Table 2-c, and the adjusted rates for occupational groups are on the last line labeled AdjO. Consider for example og[1] for administrative workers. The adjusted respiratory absence rate for the male administrative workers age 40-49 is obtained as exp(m + og[1]/100) = exp(2.85 - 0.147) = 14.91, i.e. the absence rate is about 15% lower for administrative workers, other things being equal. It is easy to see from Table 2-c that respiratory absence rates are highest in service workers (og[6] = 70.04 L%, adjusted rate = 34.8/1000) and lowest in professional workers (og[5] = - 60.49, adjusted rate = 9.44/1000), and that occupational group is an important explanatory variable for respiratory absence at this site. This follows since the ratios of the estimates og[k] to their standard errors (approximate t-tests for null hypothesis that $\theta_k = 0$) are very large for most of the estimates. A formal test of the hypothesis that there are no differences in absence rates among occupational groups (i.e. $\theta_k = 0$ for all k) is obtained as a likelihood ratio test (LRT) (see Table 3 line og). The LRT is 171.31 with 6 df indicating a large difference in occupation groups. The use of the sum to zero constraint is a

Table 3. Analysis of Deviance for Respiratory Absence Data at Site 1

	q	Deviance	IC	df	LRT	LogO
MEM	20	485.43	565.43	NA	NA	NA
Dropping Terms						
ag	17	570.07	638.07	3	84.64	40.31
og	14	656.74	712.74	6	171.31	77.42
yr	11	545.25	589.25	9	59.81	20.35
sx	19	608.40	684.40	1	122.96	64.12
Adding Terms						
ag:og	38	462.04	614.04	18	23.39	1.54
ag:yr	47	453.21	641.21	27	32.23	1.24
ag:sx	23	480.56	572.56	3	4.87	1.51
og:yr	74	425.92	721.92	54	59.51	0.93
og:sx	26	473.13	577.13	6	12.31	2.84
yr:sx	29	475.67	591.67	9	9.77	0.53

q = number of estimable parameters
IC = Deviance + q*4
LRT = Likelihood ratio test
LogO = log[(1-p)/p] where p is the p-value for LRT

mathematical and practical convenience since there is no underlying reason for selecting a particular occupation group as the reference group that can be used for all event types at all facilities. The sum to zero constraint is also used for the year explanatory variable. The ML estimates of the year effects yr[t], standard errors, RRs, and adjusted rates are given in Table 2-d. If in some situations one is interested in using a specific occupational group (say professional workers) as the reference group and a particular year (say 2000) as the reference year, then the contr.treatment contrast can be used. The details of how this is done for any contrast are described by Venables and Ripley (2002, Chapter 6.2). The methods used to implement the contrasts for each of the explanatory variables in the R function **glm**() used to fit the Poisson MEM are described in Appendix 2 – see utility function **make.h**() – and in the R documentation.

Osborn (1975) considered a multiplicative model to study the effect of several factors on vital rates using a different set of constraints. A similar approach was proposed by Mantel and Stark (1968) and Breslow and Day (1975) for the two factor case using an iterative indirect standardization technique. This technique is equivalent to fitting a MEM with two explanatory variables using Poisson regression — see Frome (1983) and Breslow and Day (1987, Chapter 4). The choice of constraints that is used results in parameters with different interpretations but does not change the value of the estimated rates for each cell in the table. This approach can be useful in the analysis of very large multidimensional tables since the fitted values for a model can be obtained using the iterative proportional fitting algorithm — see R function **loglin**(). The fitted values can then be used to calculate starting values for the linear predictor for the R function **glm**() with the desired constraints and no iteration is required since the starting values are the optimal solution.

Table 2 illustrates the results of fitting a MEM to a DOE site for one event type. The basic results are obtained from two R functions **glm**() and **predict.glm**(), and have been reformatted for convenience. In practice there could be 6 to 12 event types from the RTW absence file and the OSHA event file for each of the 13 DOE sites that participate in the IISP. In order to facilitate comparison of results, a graphical summary of parameter estimates for the MEM in Table 2 has been developed (see Figure 1). The vertical axis of the MEM plot is in log units. The dashed horizontal line is the estimated absence rate (17.28/1000) in the reference group (males age 40-49) and is shown at the top of the MEM plot. The estimated absence rates for each age group from the last line of Table 2-a (labeled AdjA) are shown as four green diamonds on the left side of Figure 1, and the corresponding names of the age groups are on the horizontal axis. The vertical distance from each diamond to the horizontal dashed line is proportional to the age effect estimates from the line Est_L% of Table 2-a, e.g. for the 30-39 age group the distance is proportional to -30.48 corresponding to an estimated rate of 12.74/1000 on the vertical axis. The gender estimates (F and M) are shown to the right of the age estimates. Note that since males age 40-49 is the reference group the effect estimate is zero and the estimated rate is on the horizontal line. The estimated effect for females is 73.34 L% (see line 2 Table 2-b) units and the vertical distance to the symbol F above the horizontal dashed line is proportional to this value. This corresponds to an estimated rate of 35.63 respiratory events/1000 person-years. The next seven symbols (see Table 1 for definition of symbols) show the adjusted rates for each of the seven occupational groups from Table 2-c. The vertical distance from each symbol to the horizontal dashed line is proportional to the value on the Est_L% row of Table 2-c. As was noted earlier, the estimated respiratory absence rate is highest in service worker and lowest in professional worker, shown as S and P in Figure 1. The estimated absence rates for years from Table 2-d are shown as squares on the right side of Figure 1 and the year names are on the horizontal axis. From Figure 1 it appears that a decrease in the respiratory absence rate occurred during the first four years followed by an increase in 1999, and then from 2000 to 2004 the rate leveled off at about 15 absences per 1000. It is clear from the estimated standard errors for the effect estimates in Table 2 that all of the covariates in Figure 1 are strong explanatory variables. This means that the corresponding RRs differ from one — e.g. the RR for female workers is 2.06 and the 95% confidence interval is
$\exp[(72.34 +/-1.96(6.45))/100] = (1.82, 2.34)$.

It is also easy to see from the main effects plot that the maximum estimated rate occurs for female service workers over age 50 in 1995, i.e. the maximum log scale rate is

$$\begin{array}{l} m + [\max(ag[i]) + \max(sx[j]) + \max(og[k]) + \max(yr[t])] / 100 \\ 2.85 + [\; 7.35 + 72.34 + 70.04 + 40.79 \;] / 100 = 4.755, \end{array}$$

and the maximum estimated respiratory event rate is exp(4.755) = 116/1000. The same result can be obtained graphically by adding the vertical distance from the horizontal line for female (F), service workers (S) and the year 1995 onto the triangle for the over 50 age group. The lowest estimated rate occurs for professional male workers under age 30 in 2004, i.e. the minimum log scale rate is

$$\begin{array}{l} m + [\min(ag[i]) + \min(sx[i]) + \min(og[k]) + \min(yr[t])] / 100 \\ 2.85 + [\; -124.67 + 0 + -60.47 + -21.02 \;] / 100 = 0.788, \end{array}$$

and the estimated respiratory event rate is 2.2/1000.

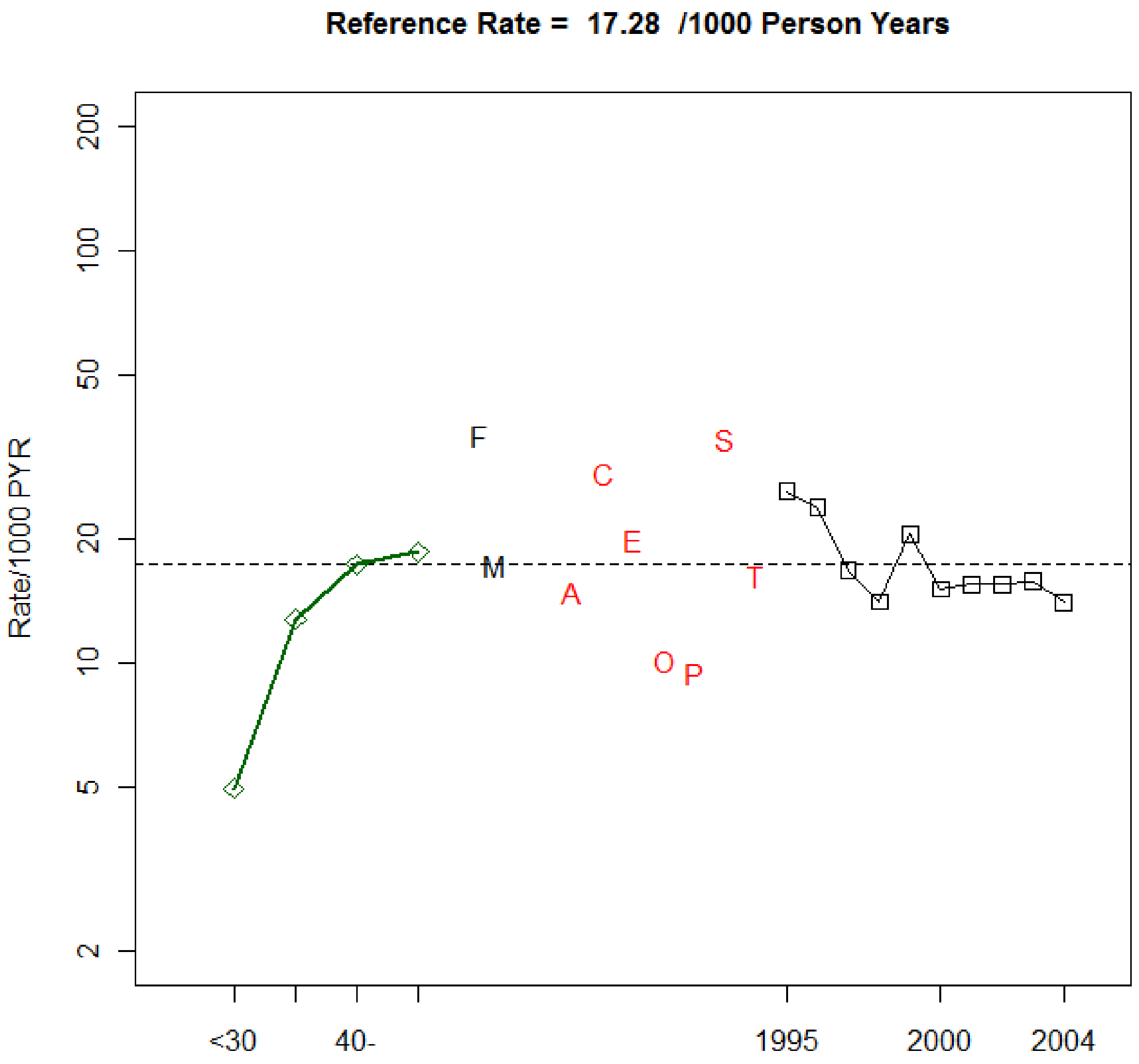

Figure 1: Estimated Rates for Respiratory Absences at Site 1

The Poisson analysis of deviance table is obtained by evaluating the deviance $D(\hat{\beta})$ (equation 4) for each of a series of models (see Table 3). In Table 3 column 1 describes the model, column 2 the number of parameters, and column 3 is the value of the deviance. It also includes the IC (equation 5) for each model, and a likelihood ratio test (LRT) statistic that is used to evaluate the importance of each term in the MEM and each of the possible two factor interactions. The first row in Table 3 (labeled MEM) lists the number of parameters, the deviance and the IC for the MEM. The next four rows are obtained by dropping each of the explanatory variables from the MEM. For example, row 3 (labeled og) shows the results of dropping the og term from the MEM. The reduced model has 14 parameters, the deviance is 656.74, and the IC is 656.74 + 4*14 = 712.74. The likelihood ratio test for the null hypothesis that there is no difference in absence rates for occupational groups is 171.31 (see row 3 column 6) with 6 df. This is compared to the chi-square distribution with 6 df. The last column in Table 3 gives the value of Log0 = log[(1-p)/p], where p is the probability that a chi-square statistic with 6 df is greater than 171.31. This ad-hoc statistic is the log odds of observing a likelihood ratio test that exceeds the calculated value if the null hypothesis is true. For this example it is the log odds of observing a chi-square statistic with 6 df that exceeds 171.31 if there is no difference in respiratory absence rates for the occupational groups (note that the value of p is 2.372 x 10^{-24}). This shows that occupational group is a very strong explanatory variable for respiratory absence rates at this site, as are all of the terms in the MEM (as indicated by the Log0 values in column 7). This is also evident from the IC values in column 4, which are also considerably larger than the IC value for the MEM. The next six rows show the effect of adding each of the possible two factor interactions to the MEM. For example, adding the og:yr interaction decreases the deviance by (485.43 – 425.92) = 59.51. This is a LRT of the null hypothesis that the og:yr interactions are all zero. The LRT statistic follows a chi-square distribution with 54 df when the null hypothesis is true resulting in a p-value of 0.282 and Log0 = 0.93. Rejecting the null hypothesis for $p < 0.05$ (or for $p < 0.01$) is equivalent to rejecting the null hypothesis for Log0 > 2.94 (or for Log0 > 4.60).

The IC in column 4 is used to compare the goodness of fit of each of the models with an interaction term added to the MEM. Adding a term to a model will always decrease the deviance so adding 4*q to the deviance is a penalty for increasing the complexity of the model. In this example the IC for the MEM (565.43) is the smallest IC value in the table. This shows that all the explanatory variables in the MEM are important and that none of the two factor interactions are important as is also indicated by the Log0 values in column 7 for the likelihood ratio test statistics in column 6 of Table 3. The value of the IC is used to identify interaction terms that may be important. If the IC for one or more of the interaction terms is less than the IC for the MEM then further analysis is considered (see Discussion).

3.2 EXAMPLE 2

To further illustrate the use of Poisson regression methods, analysis of respiratory events at a second site described in the Worker Health Summary is presented. The results from fitting the MEM to Site 2 are shown in Figure 2 and Tables 4 and 5. The scale on the vertical axis in Figure 2 is the same as in Figure 1.

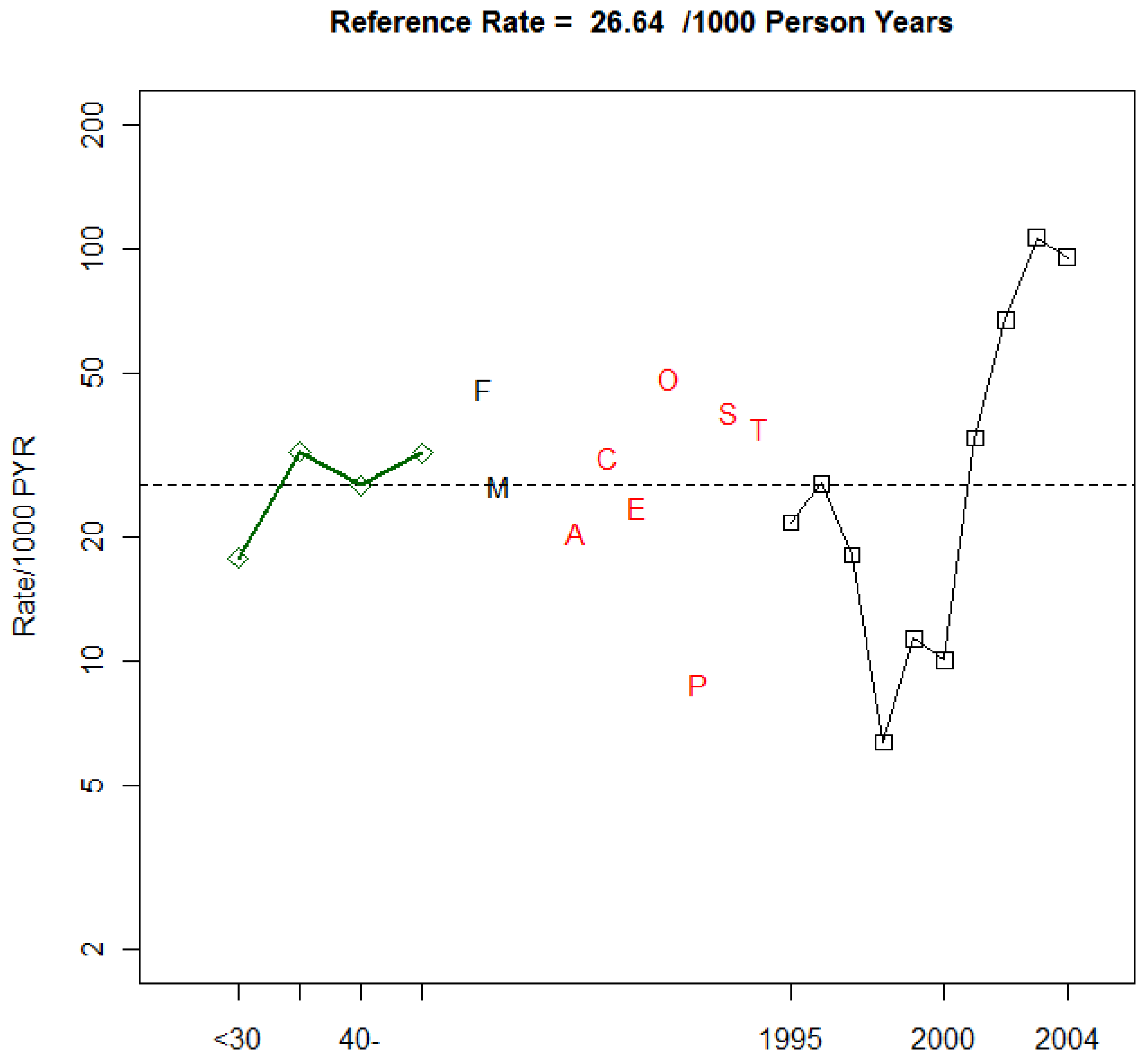

Figure 2: Estimated Rates for Respiratory Absences at Site 2

The estimated rates are about 50% higher at Site 2 than Site 1, and the RR for women is increased at both sites. The estimates for occupational groups A, C, E, P, and S follow a similar pattern at both sites, and the rates for O and T are relatively higher at Site 2. The effects for year are also clearly different over the ten-year period. Site 1 shows a decreasing trend over the first four years with a slight increase in 1999 followed by a level response over the last five years. Site 2 also shows an initial decrease over the first four years followed by a strong increase in the respiratory absence rates (about 35L% per year) over the last six years [(127-(-85))/6= 35]. These differences are more clearly determined from comparing the parameter and rate estimates in Tables 2 and 4.

Table 4. Poisson Regression Results for Respiratory Absences at Site 2

Minimal Model: Deviance = 2192.86, n = 554
Rate(unadjusted)= 47.31 resp events/1000, m = 3.857

Main Effects Model: log(rate)= m + ag + sx + og + yr

Deviance = 635.685, df = 534, q = 20, IC = 715.68, R2 = 69.9%
Dispersion: phi.p = 1.26, Pearson X2 = 672.2, Score Ta = 5.71
Rate(adjusted)= exp(m)= 26.64 resp events/1000, m = 3.282

a) Age Groups: ag[] Estimates

	<30	**30-**	**40-**	**50+**
Est_L%	-40.42	18.59	0.00	18.02
SE_L%	11.38	6.86	0.00	6.26
RR	0.67	1.20	1.00	1.20
AdjA	17.78	32.08	26.64	31.90

b) Gender: sx[] Estimates

	F	**M**
Est_L%	54.45	0.00
SE_L%	5.90	0.00
RR	1.72	1.00
AdjG	45.92	26.64

c) Occupational Groups: og[] Estimates

	A	**C**	**E**	**O**	**P**	**S**	**T**
Est_L%	-14.77	50.74	13.90	-53.52	-60.47	70.04	-5.92
SE_L%	7.91	11.15	19.09	27.42	7.30	10.93	8.72
RR	0.86	1.66	1.15	0.59	0.55	2.01	0.94
AdjO	14.91	28.71	19.86	10.12	9.44	34.82	16.29

d) Year: yr[] Estimates

	1995	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
Est_L%	-20.74	1.13	-38.72	-143.30	-85.21	-97.44	26.39	92.42	138.36	127.11
SE_L%	10.37	9.36	11.24	19.47	14.85	15.83	8.82	6.80	5.99	6.15
RR	0.81	1.01	0.68	0.24	0.43	0.38	1.30	2.52	3.99	3.56
AdjY	21.65	26.94	18.09	6.36	11.36	10.05	34.69	67.13	106.27	94.97

**The Reference Group for Adjusted Rates is Males Age 40-49*

The estimates of the dispersion parameter and score statistic T_a in Table 4 indicate over-dispersion. This could be due to lack-of-fit of the MEM, extra-Poisson variation, or both. The next step in the analysis would be to use model checking procedures to determine if there are a few outlying y values and/or model inadequacies that are having an undue influence on the test for over-dispersion.

The results in the analysis of deviance table from Site 2 are shown in Table 5. Log0 values greater than 2.94 indicate that several of the two-level interactions are significant at the 0.05 level of significance. This matter will be considered in Section 3.3.

Table 5: Analysis of Deviance for Respiratory Absence Data at Site 2

	q	Deviance	IC	df	LRT	LogO
MEM	20	635.68	715.68	NA	NA	NA
Dropping Terms						
ag	17	673.61	741.61	3	37.92	17.34
og	14	1,063.66	1,119.66	6	427.98	203.94
yr	11	1,638.38	1,682.38	9	1,002.69	482.03
sx	19	717.39	793.39	1	81.71	43.29
Adding Terms						
ag:og	38	554.89	706.89	18	80.80	21.20
ag:yr	47	587.84	775.84	27	47.85	4.82
ag:sx	23	628.26	720.26	3	7.42	2.76
og:yr	74	546.61	842.61	54	89.08	6.28
og:sx	26	606.05	710.05	6	29.63	9.98
yr:sx	29	625.06	741.06	9	10.62	0.83

q = number of estimable parameters
IC = Deviance + q*4
LRT = Likelihood ratio test
LogO = log[(1-p)/p] where p is the p-value for LRT

3.3 OVER-DISPERSION

An informal procedure that is used to check for systematic departures from the Poisson MEM is based on four regression diagnostic plots (see Methods). To illustrate this procedure, we use the results from the MEM for Example 2. The score test $T_a = 5.71$ ($p < 10^{-6}$) and estimates of the dispersion parameter greater than one seen in Table 4 and the Log0 values greater than 2.94 for two way interaction effects in Table 5 indicate lack of fit of the MEM, heterogeneity of variance, or both.

Figure 3 contains the four regression diagnostic plots. A plot of the SDRs against the fitted rates from the MEM is shown in Figure 3a. The R function **supsmu**( ) was used to calculate the solid line, and the two dashed lines correspond to the 0.005 and 0.995 quantities of the standard normal distribution, i.e. if the SDRs are approximately N(0,1) about 99% of these residuals should be between the dashed lines. The seven SDRs outside the 99% limits are marked with a red x. For model checking purposes, a normal Q-Q plot is used to identify extreme values which would appear in the upper right and/or lower left portion of the plot (see Figure 3b). The solid line in Figure 3b corresponds to the standard normal distribution. The SDR-Leverage plot in Figure 3c identifies four points with both $h_i > 2$ (i.e. to the right of the solid vertical line) and SDRs that are outside the horizontal dashed lines at +/- 2.58. A plot of the absolute value of the SDRs against the fitted values (see Figure 3d) gives an informal check of on the adequacy of the assumed variance function. The null pattern will not show a trend, and smoothing (shown by the solid line) is used to identify a possible pattern, in this case a positive trend.

The results for Example 2 in Tables 4 and 5 and Figure 3 indicate that MEM did not fit well. The score static $T_a = 5.71$ and the apparent trend in Figure 3d both indicate over-dispersion, and the other three diagnostic plots do not indicate that “outliers” are a problem. There are several possible explanations – lack-of-fit of the MEM, extra-Poisson variation or both. The IC values in Table 5 suggest that adding the ag:og and/or the og:sx and/or the og:yr interaction terms may result in a “better” model. It is also possible that higher order interactions may be important, e.g. the ag:og:sx three way interaction.

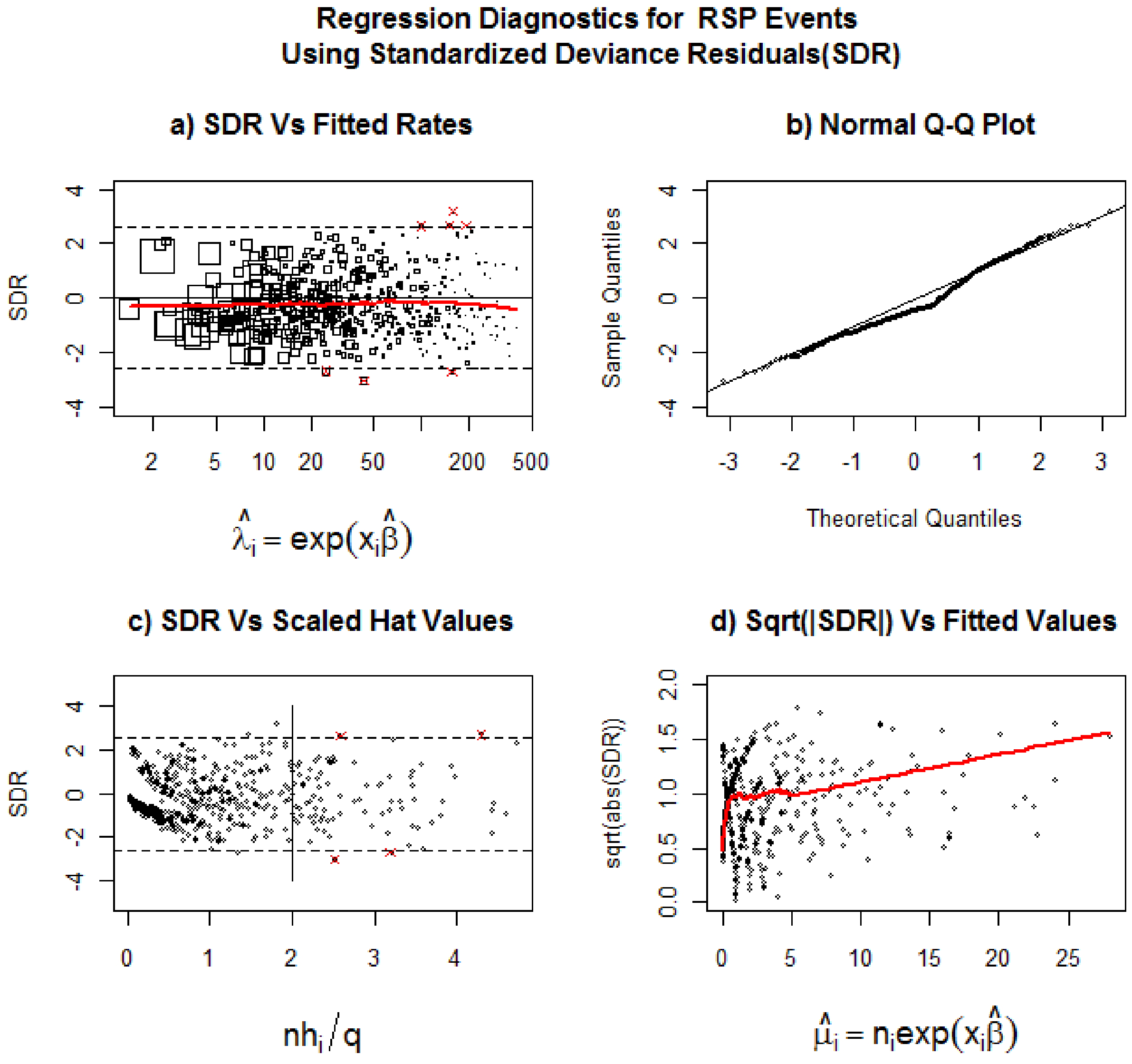

Figure 3: Regression Diagnostics for Respiratory Absences at Site 2

Venables and Ripley (2002, Chapter 6) describe various approaches to the analysis of Poisson data that can be used to evaluate the need for additional terms in the model. The MEM serves as the starting point, and the function requires a list defining the most complex and the simplest models to be considered. The value k = 4 is used in the IC (see equation 5) and an additional argument “direction” is used to specify that the process should add terms or remove terms as needed. This model selection procedure evaluates the effect of adding or dropping terms from the model subject to the “marginality restriction”, which imposes certain constraints on the underlying model matrix. For example, if a two factor interaction, say ag:og is in a model, then both main effects ag and og must be in the model. The algorithm evaluates the IC for each model and returns the model with the smallest IC value. The call to **stepAIC**() appears as follows:

```
stepAIC(MEM, scope=list(upper = ˜ (ag+og+yr+sx) ^3, lower = ˜ 1, directions = “both”, k=4)
```

The most complex model considered contains all possible three factor interactions (and due to marginality constraints all possible 2 factor interactions and all four main effects), and the simplest model is the one parameter minimal model. The resulting “best” model for the second example is the MEM with two interaction terms ag:og and og:sx. The deviance for this model is 527.44 with 510 df, the IC is 527.44 + 4*44 = 703.44, and the score test for this model is $T_a = 1.263$ ($p = 0.103$) indicating that the apparent over-dispersion in this example could be attributed to lack-of-fit of MEM. The **effects** package in R (Fox, 2003) can be used to obtain a graphical display of terms in a complex GLM, i.e. in this situation, a model that contains interactions. The effect display for the ag:og interaction is shown in Figure 4. The estimated rates are on the vertical scale in log units as in the MEM plot in Figure 2. The effect display in Figure 4 provides a visual representation of the interaction effects (see the help file for the **effects** package for details).

An alternative way to evaluate a two factor interaction that may be of interest is to obtain a two-way marginal table of observed y_{ij} and expected $\hat{y}_{ij}$ values, where the expected values are obtained from the MEM. The statistic $(y_{ij} - \hat{y}_{ij}) / (\hat{y}_{ij})^{1/2}$ is calculated for each cell in the marginal table to identify the source of the interaction. We also calculate $\log(r_{ij} / \hat{\lambda}_{ij})*100$ to obtain an approximate estimate of the interaction term for each cell in the table, where $r_{ij} = y_{ij}/n_{ij}$ is the observed marginal rate.

The most common approach that is used to deal with over-dispersion is to assume that there is extra-Poisson variation as described by the variance function in equation (6) with b=1, i.e. QL/M1 regression. For Example 2 we proceed under the assumption that the MEM is acceptable and that the over-dispersion is due to extra-Poisson variation. From the Table 4 Dispersion line, $\tilde{\varphi} = 1.26$. The parameter estimates in Table 4 remain unchanged and the standard errors are multiplied by $\sqrt{\tilde{\varphi}} = 1.122$. The likelihood ratio tests in Table 5 are converted to F tests and the adjusted IC (ICa) values based on equation (5) with $\varphi = \tilde{\varphi}$ are calculated using the **stepAIC**() function with $k = 4\tilde{\varphi}$. The adjusted analysis of deviance table that appears in Table 6 shows that the smallest ICa value occurs for the MEM.

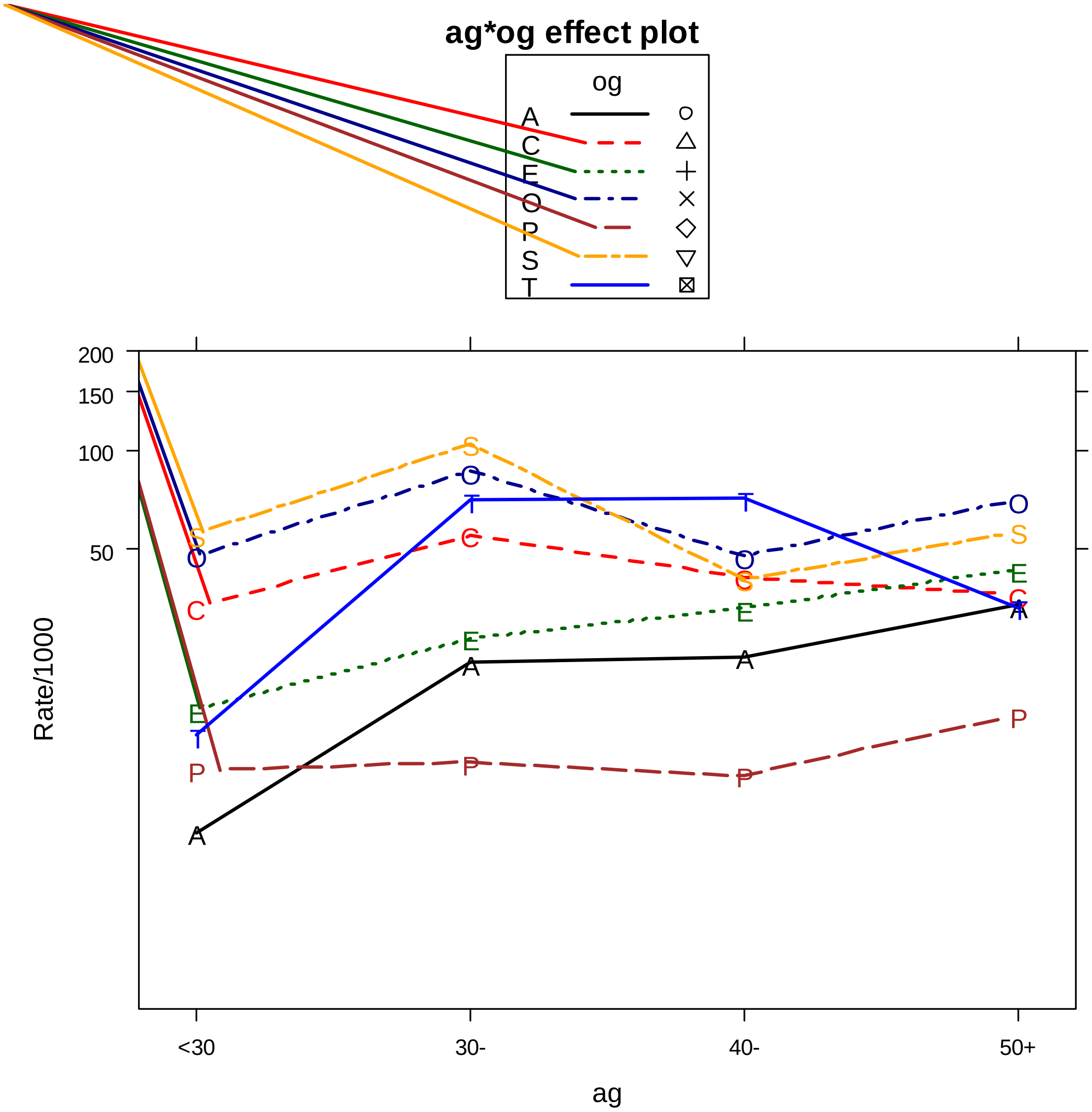

Figure 4: Effect Plot for Occupational Group Interaction with Age Group

To summarize the over-dispersion investigation based on the respiratory absences at Site 2, the score test for over-dispersion T_a is highly significant. There are, however, two possible explanations. The first explanation attributes the over-dispersion to lack-of-fit of the MEM due to the failure to include two interaction terms. The second explanation is based on QL/M1 analysis and attributes the over-dispersion to extra-Poisson variation, and concludes that the MEM (with adjusted test statistic) is a reasonable model. In the IISP the MEM is of primary interest since it facilitates the visual comparison of the event rates for a number of different event types from the RTW and OSHA files at each of the participating DOE sites. Therefore, in the context of the IISP the second explanation is preferred since the inclusion of interaction terms does not result in substantial improvement over the MEM, i.e. the MEM is the preferred model unless there is strong evidence that it is not acceptable.

Table 6. Analysis of Deviance for Respiratory Absences at Site 2 Adjusted for Over-dispersion

	q	Deviance	IC	df	LRT	LogO
MEM	20	635.68	736.39	NA	NA	NA
Dropping Terms						
ag	17	673.61	759.20	3	10.04	13.17
og	14	1,063.66	1,134.15	6	56.67	122.93
yr	11	1,638.38	1,693.76	9	88.51	228.42
sx	19	717.39	813.06	1	64.91	32.90
Adding Terms						
ag:og	38	554.89	746.22	18	3.57	13.66
ag:yr	47	587.84	824.48	27	1.41	2.38
ag:sx	23	628.26	744.07	3	1.97	2.01
og:yr	74	546.61	919.20	54	1.31	2.52
og:sx	26	606.05	736.96	6	3.92	7.19
yr:sx	29	625.06	771.08	9	0.94	0.03

ICa = Deviance + q*4*1.259
F= F Test
LogO= log[(1-p)/p] where p is the p-value for LRT

In a general context, if the initial analysis of the data indicates over-dispersion that could be explained by the addition of interaction terms to the MEM, then the results from the QL/M1 analysis can be extended to evaluate the importance of interaction terms, in addition to the usual informal model checking procedures. For the QL/M1 analysis the parameter estimates are equal to the ML estimates under the Poisson assumption. To determine if adding interaction or deleting main effect terms could result in a "better" model, one can use the IC based on the Poisson log-likelihood and increase the penalty for unnecessary terms by using the moment estimate $\tilde{\varphi}$ (calculated for the MEM) in the penalty term (see equation 5) calculated with the **stepAIC**() function with k = $4\tilde{\varphi}$. If the adjusted IC or corresponding F test indicates that one or more interaction terms are important then further analysis can be considered. This would occur, for example, if one of the main effects in the important interaction term is a weak explanatory variable as indicated by a small ICa value and a non-significant F test (i.e. a LogO < 2.94).

4. DISCUSSION

Poisson regression models are widely used in the analysis of count data in many areas of data analysis as described in the Introduction. In some situations there is over-dispersion and, when the primary interest is in the explanatory variables of the regression function, the analysis can be adjusted using the QL/M approach with a linear or quadratic variance function. The counts for each observational unit are assumed to follow the Poisson distribution with means that vary as described by a mixing distribution (e.g. a gamma distribution). Using QL/M requires only that the first two moments of the distribution of the counts be specified. Dean (1994) describes the use of quasi-likelihood to estimate the regression parameters (β) and an additional estimating equation for the dispersion parameter δ. Five methods for estimation of δ are presented and the advantages of this approach to handling over-dispersion are discussed. Most of these methods are easy to implement with only minor adjustments to standard software using the IWLS approach (see Appendix 2). Zeileis et al. (2008) describe the computational aspects of fitting these models from an estimating function point of view and describe R model fitting functions and associated methods for diagnostics and inference. The properties of QL/M methods have been studied using asymptotic theory and Monte Carlo simulations--- see e.g. Lawless (1987), Moore (1986), Breslow (1990), Dean (1992), and Dean (1994). Further extensions of this type of modeling by specifying low-order conditional moments of the dependent variable are described by Cameron and Trivedi (1998, Chapter 12). Hilbe (2008, Chapter 4) provides a detailed discussion of over-dispersion and addition issues that may be useful when identifying and dealing with over-dispersion. An alternative to QL/M when over-dispersion occurs is to replace the Poisson distribution with a two-parameter count distribution. Puig and Valero (2006) characterize all two-parameter count distributions that are partially closed under addition and for which the sample mean is the ML estimate of the population mean. Mixed Poisson models such as the NB satisfy this property and can be used to account for over-dispersion using the full ML approach. Kim and Kriebel (2009) have proposed the use of NB regression for the analysis of health events surveillance data that may be over-dispersed. In some situations with over-dispersion the ML estimate of the dispersion parameter may be zero and the NB estimates reduce to the Poisson estimates. This is illustrated by Lawless (1987) using the ship damage data from McCullagh and Nelder (1989, Chapter 6), and it is noted that in this example that the QL/M1, QL/M2, and NB regression yield quite different estimates of the standard errors for the regression parameters. There is no formal statistical way to choose between the two QL methods as far as we know. Consequently, we use QL/M1 in situations where adjusting for over-dispersion is indicated.

A second problem that sometimes occurs in the analysis of counts is the occurrence of “too many zeros.” That is, in addition to or instead of extra-Poisson variation or lack-of-fit of the regression model being the source of over-dispersion, there may be too many zeros relative to a Poisson or NB regression model. The models that have been proposed to deal with this are described in Appendix 1, and are generally referred to as “zero-altered” models. The application of these models to the IISP data would require an assumption that a health event is described by a model in which it is expected that a zero count will occur more often than predicted based on a Poisson or NB regression model and the occurrence of zero counts requires special consideration. There is no apparent logical basis for this situation to occur for the health events and related data structures of interest to the IISP program.

5. ACKNOWLEDGEMENTS

This document describes activities performed with support from the Office of Domestic and International Health Studies, Office of Health and Safety of the U. S. Department of Energy (DOE), and was performed at the Center for Epidemiologic Research at the Oak Ridge Institute for Science and Education (ORISE), which is managed by Oak Ridge Associated Universities under contract number DE-AC05-06OR23100. ORISE is a DOE institute focusing on scientific initiatives to research health risks from occupational hazards, assess environmental cleanup, respond to radiation medical emergencies, support national security and emergency preparedness, and educate the next generation of scientists. Additional support was provided by Computer Science and Mathematics Division at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725. The authors thank B. Jolene Jones for assistance in manuscript preparation, and George Ostrouchov and Paul Wambach for comments on the material in this report and the related website. This work has been authored by a contractor of the U.S. Government. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this work, or to allow others to do so for U. S. Government purposes.

REFERENCES

Atkinson, A.C. 1981. Likelihood ratios, posterior odds, and information criteria. *Journal of Econometrics* **16(1)**: 15-20.

Bailer, A.J. Reed, L.D. and Stayner, L.T. 1997. Modeling fatal injury rates using poisson regression: A case study of workers in agriculture, forestry, and fishing. *Journal of Safety Research* **28(3)**: 177-86.

Baker, R.J. and Nelder, J.A.1978. *Generalized linear interactive modeling (GLIM), Release 3*. Oxford: Numerical Algorithms Group.

Bishop, YM, Feinberg, SE, & Holland, PW ,1975 . Discrete Multivariate Analysis: Theory and Practice, The MIT Press , Cambridge.

Bortkiewicz, L von, 1898, Das Gesetz der kleinen Zahlen [The law of small numbers] Leipzig, Germany: B.G. Teubner.

Breslow, N.E. 1984. Extra-Poisson variation in log-linear models. *Applied Statistics* **33(1):** 38-44.

Breslow, N. 1990. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association* **85(410)**: 565-71.

Breslow, N.E. and Day, N.E. 1987. *Statistical methods in cancer research, volume II: the design and analysis of cohort studies.* Number Scientic Publication 82. International Agency for Research on Cancer, Lyon.

Brown MB and Fuchs C, 1983. On maximum likelihood estimation in sparse contingency tables. *Computat. Statist. Data Anal* **1**:3-15.

Cameron, A.C. and Trivedi, P.K. 1986. Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of applied Economics* **1**:29-53.

Cameron, AC and Trivedi, PK. 1998. *Regression Analysis of count data*. Econometric society Monograph No. 30, Cambridge University Press.

Carrivick PJW, Lee AH, Kelvin, and Yauc KKW. 2003. Zero-inflated Poisson modeling to evaluate occupational safety interventions. Safety Science **41**:53–63.

Charnes A, Frome EL, and Yu PL. 1976. The equivalence of generalized least squares and maximum likelihood estimation in the exponential family. *Journal of the American Statistical Association* **71**:169-72.

Clayton D, and Hills M. 1993. *Statistical models in epidemiology.* Oxford University Press, New York.

Cochran, WG. (1940) The analysis of variance when experimental errors follow the Poisson or binomial law. *Annals of Mathematical Statistics* **11**: 335-347.

Dean CB, and Lawless JF. 1989. Tests for detecting over-dispersion in Poisson regression models. *Journal of the American Statistical Association* **84:** 467-71.

Dean CB. 1992. Testing for over dispersion in Poisson and binomial regression models. *Journal of the American Statistical Association* **87(418):** 451-57.

Dean CB. 1994. Modified pseudo-likelihood estimator of the over-dispersion parameter in Poisson mixture models. *Journal of Applied Statistics* **21(6)**: 523-32.

Erlander S, Gustavsson J, and Svensson A. 1972. On asymptotic simulations confidence regions for regression planes in a Poisson model. *Longman Group Limited* **40(2)**: 111-22.

Fox J. 2003. Effect displays in R for generalised linear models. *Journal of Statistical Software* **8(15):** 1-18.

Fridstrom L, Ifver J, Ingebrigtsen S, Kulmala R, and Thomson LK. 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis and Prevention* **27(1)**: 1-20.

Frome EL, Kutner MK, and Beauchamp JJ. 1973. Regression analysis of Poisson distributed data. *Journal of the American Statistical Association* **68**: 935-40.

Frome EL, and Walton C M. 1975. *A method for assessing the impact of the energy crisis on highway accidents in Texas.* Council for Advanced Transportation Studies, Austin.

Frome EL. 1983. The analysis of rates using Poisson regression models. *Biometrics* **39(3)**: 665-74.

Frome EL. 1984. Response to Nelder's reaction on Poisson rate analysis. *Biometrics* **40**: 1160-62.

Frome E, and Checkoway H. 1985. Use of Poisson regression models in estimating incidence rates and ratios. *American Journal of Epidemiology* **121(2):** 309-23.

Frome EL, Cragle DL, Watkins JP, Wing S, Shy C, Tankersley WG, and West CM. 1997. A mortality study of employees of the nuclear industry in Oak Ridge, Tennessee. *Radiation Research* **148**: 64-80.

Green PJ. 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *The Journal of the Royal Statistical Society . Series B (Methodological).* **46(2):** 149-92.

Gustavsson J. and Svensson A. 1976. A Poisson regression model applied to classes of road accidents with small frequencies. *Scandinavian Journal of Statistics* **3(2)**:49-60.

Haight FA. 1967. *Handbook of the Poisson Distribution*. John Wiley & Sons, New York.

Haslett S. 1990. Degrees of freedom and parameter estimability in hierarchical models for sparse complete contingency tables. *Computational Statistics & Data Analysis* **9**:179- 195.

Heinzl H, and Mittlbock M. 2003. *Pseudo R-squared measures for Poisson regression models with over- or underdispersion*. Computational Statistics & Data Analysis **44**, www.ElsevierMathematics.com

Hilbe, JM. 2008. *Negative Binomial Regression.* New York, Cambridge University Press.

Jackman S. 2008. pscl: *Classes and Methods for R Developed in the Political Science Computational Laboratory*, Stanford University. Department of Political Science, Stanford University, Stanford, California. R package version 0.95, URL http://CRAN.R-project.org/ package=pscl.

Jorgenson, D. W., (1961), Multiple Regression Analysis of a Poisson Process, Journal of the American Statistical Association, **56**: 235-245.

Karra, V., (2005), Analysis of Non-fatal and Fatal Injury Rates for Mine Operator and Conractor Employees and the Influence of Work Location, **36**: 413-421.

Khan,A. Ullah, S. and Nitz,J. 2011 , Statistical modelling of falls count data with excess zeros. Injury Prevention **17:4** 266-270.

Koch, G. G., Atkinson, S. S. and Stokes, M E., (1984), Poisson Regression, in: Encyclopedia of Statistical Sciences, eds Kotz, S., Johnson, L. L. and Read, A, New York: J. Wiley & Sons, pages 32-40.

Laaksonen M, Martikainen P, Rahkonen O, and Lahelma E. 2008. Explanations for gender differences in sickness absence: evidence from middle-aged municipal employees from Finland. *Occup Environ Med* **65**:325–330. doi:10.1136/oem.2007.033910.

Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics **34(1):**1-14.

Li, G., Shahpar, C., Grabowski, J.G., and Baker, S.P., (2001), Secular Trends of Motor Vehicle Mortality in the United States, 1910-1994, Accident Analysis and Prevention, 33, 423-432.Lawless JF. 1987. Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15(3):** 209-25.

Lord D, Washington SP, and Ivan NJ. 2005 b. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention* **37(1)**: 35-46.

Lord D, Washington SP, and Ivan JN. 2007. Further notes on the application of zero inflated models in highway safety. *Accident Analysis & Prevention* **39 (1):** 53-57.

Maindonald J and Braun J. *Data Analysis and Graphics Using R*. Cambridge University Press, Cambridge, 2nd edition, 2007. ISBN 978-0-521-86116-8.

Mallick S, and Mukherjee K. 1996. An Empirical Study for Mines Safety Management through Analysis on Potential for Accident Reduction. *Omega, The International Journal of Management Sci*ence, **24(5)**: 539-50.

Mantel, N. and Stark, C. R., (1968), Computation of Indirect Adjusted Rates in the Presence of Confounding, Biometrics, **24**: 997-1005.

McCullagh P, and Nelder JA. 1989. *Generalized Linear Models*. Chapman and Hall, New York.

Melchlor M, Krieger N, Kawachi I, Berkman LF, Niedhammer I, and Goldberg M. 2005. Work factors and occupational class disparities in sickness absence: findings from the GAZEL cohort study. *American Journal of Public Health* **95(7)**: 1206-1212. doi: 10.2105/AJPH.2004.048835.

Michener R, and Tighe C. 1992. A Poisson regression model of highway fatalities. The *American Economic Review. Papers and Proceedings of the Fourth Annual Meeting of the American Economic Association,* **82(2)**, 452-57.

Mohr DL, and Clemmer DI. 1989. Evaluation of an occupational injury intervention in the petroleum drilling industry. *Accidental Analysis & Prevention* **21(3)**: 263-71.

Moore, L. M. and Beckman, R. J., (1988), Approximate One-Sided Tolerance Bounds on the Number of Failures Using Poisson Regression, Technometrics, **30**: 283-290.

Mullahy J (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics* **33**: 341-365.

National Center for Health Statistics (NCHS) and the Centers for Medicare & Medicaid Services (CMS) (2011): International Classification of Diseases, 9th Revision, Clinical Modification" (ICD-9-CM), Sixth Edition (CD-ROM). Washington, DC: U.S. Government Printing Office.

Nelder JA, and Wedderburn RWM. 1972. Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135(3)**: 370-84.

Ostrouchouv G, and Frome EL. 1993. A model search procedure for hierarchical models. *Computational Statistics and Data Analysis,* **15(3)**: 285-96.

Puiz P. and Valero, J., 2006, Count Data Distributions: Some Characterizations with Applications, *Journal of the American Statistical Association*, **101:473**, 332-340.

Pierce DA, Preston DL, Vaeth M, and Mabuchi K. 1996. Studies of the Mortality of Atomic Bomb Survivors, *Radiation Research*, Report 12, Part I, Cancer: 1950-1990, **146(1):** 1-27.

Preece DA, Ross GJS, and Kirby PJ. 1988. Bortkewitsch's horse-kicks and the generalized linear model. *Journal of the Royal Statistical Society. Series D (The Statistician)* **37(3):** 313-18.

Pregibon D. 1981. Logistic regression diagnostics. *The Annals of Statistics* **9(4)**: 705-24.

Preston DL, Lubin JH, Pierce DA, and McConney ME. 1993. *Epicure User's Guide*, Technical Report, Hiresoft International Corporation, Seattle.

Quine MP, and Seneta E. 1987. Bortkiewicz's data and the law of small numbers. *International Statistical Review* **55(2)**: 173-81.

R Development Core Team.2011. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, www.R-project.org.

Richardson DB, and Loomis D. 2004. The impact of exposure categorisation for grouped analyses of cohort data. *Occupational and Environmental Medicine* **61(11):** 930-35.

Richardson DB, Loomis D, Bena J, and Bailer AJ. 2004. Fatal occupational injury rates in southern and non-southern states, by race and Hispanic ethnicity. *American Journal of Public Health* **94:** 1756-761.

Ridout M, Hinde J, and Demetrio CG. 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57**: 219-23.

Rose CE, Martin SW, Wannemuehler KA, and Plikaytis BD. 2006. On the use of zero-inflated and hurdle modes for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics,* **16**: 463–481. DOI: 10.1080/10543400600719384.

Rostguard K. 2008. Methods for stratification of person-time and events – a prerequisite for Poisson regression and SIR estimation. *Epidemiologic Perspectives & Innovations* **5**: 1-16.

Smitha, MW, Kirk KA, Oestenstad KR, Brown KC, Lee SD. 2001. Effect of state workplace safety laws on occupational injury rates. *Journal of Occupational and Environmental Medicine*, **43(12)**: 1001-1010.

Strader C, and Richter B. 2007. U.S. Department of Energy Illness and Injury Surveillance Program Worker Health Summary, 1995-2004, 07-OEWH- 1073, http://www.hss.energy.gov/HealthSafety/IIPP/hservices/epi_surv.html.

Slymen, D., Ayala, G., Arredondo, E., Elder, J., 2006, A demonstration of modeling count data with an application to physical activity. Epidemiologic Perspectives and Innovations **3**: 1-9.

Tornqvist L, Vartia P, and Vartia YO. 1985. How should relative changes be measured? *The American Statistician* **39(1)**: 43-46.

Venables WN and Ripley BD. *Modern Applied Statistics with S. Fourth Edition*. Springer, New York, 2002. ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4/.

Ullah S, Finch C, and Day L. 2010. Statistical modelling for falls count data. *Accident Analysis and Prevention* **42**:384–392.

Wing S, Shy CM, Wood JL, Wolf S, Cragle DL, and Frome EL. 1991. Mortality among workers at Oak Ridge National Laboratory: Evidence of radiation effects in follow-up through 1984. *Journal of the American Medical Association* **265(11)**: 1397-1402.

Zeileis A, Kleiber C, and Jackman S. 2008. Regression models for count data in R. *Journal of Statistical Software* **27:** 1-12.

APPENDIX 1. ZERO-ALTERED MODELS

In the Methods and Discussion sections the issue of over-dispersion in Poisson regression analysis was examined and the QL/M and NB regression were considered to adjust the analysis for extra Poisson variation. Another manifestation of over-dispersion that may occur when Poisson regression models are used is an excess number of zero counts, i.e., the observed number of zero events is much greater than the expected number of zeros based on the Poisson regression. In some cases, the use of the NB distribution will resolve the situation. When this is not the case two zero-altered (ZA) models – hurdle and zero inflated (ZI) models – have been proposed to deal with the count data having an excess of zero counts. Hilbe (2008, Chapter 8) and Cameron and Trivedi (1998, Chapter 4) described these ZA models and their relationship to other models for count data. Application of these methods to injury count data have been described by Carrivick et al (2003), Lord et al (2005), Slymen et al (2006), Ullah et al (2010), and Kahn et al (2011). Rose et al (2006) consider the use of these models in public health studies with the analysis of vaccine adverse event counts as a motivating example.

In generic terms, there are two types of ZA models. The first type of ZA model is based on the assumption that the process determining that an event occurs may differ from the process that determines how many events occur when there is at least one event. These conditional models are also referred to as "two-part" models or "hurdle" models. The second type of model is a mixture of a degenerate mass at zero and a Poisson (or NB) distribution, and is referred to as a zero inflated (ZI) model or a zero inflated Poisson model (ZIP) when the Poisson distribution is utilized. The ZI model is also described as a "dual state process" model and differs from the hurdle model in that there may be either "structural" zeros or sampling zeros while hurdle models have only sampling zeros. For illustrative purposes we use a propagation experiment in which the number of roots (y) produced by a plant cutting (micro-propagated shoots of an apple tree) during a period of time is the count variable of interest – see Ridout et al (2001) for additional details and a numerical example. This example is used for discussion since the distinction between the two types of models would be based on realistic biological assumption, i.e. it does not make sense to fit hurdle and ZI models to the same data. If only sampling zeros are possible, then a hurdle model is indicated. The entomologist may believe that the mechanism that determines how many roots occur on cuttings differs from the mechanism that determines whether or not a cutting can root at all, which would indicate a hurdle model. For this two part model, there is a probability π that a shoot fails to root, and for shoots that do root the distributions of the number of roots is described by a zero-truncated discrete distribution. If the Poisson distribution is used for the positive counts the distribution of Y is defined as follows:

$$\Pr[Y_i = y_i] = \begin{cases} \pi_i & y_i = 0 \\ (1-\pi_i)\mathrm{f}(y_i) \ / \ [1-\mathrm{f}(0)] & y_i > 0\,, \end{cases} \qquad (9)$$

where $f(y) = \exp(-\lambda)\ \lambda^y\ / y!$.

The parameters λ_i and π_i may depend on vectors of covariates x_i and z_i, respectively. The first part of the model is described by a binomial probability that determines whether a count variable has a zero or positive value, e.g. a binomial logit GLM is the most intuitive specification. If the value is positive, the counts are described by a zero-truncated Poisson distribution. This "hurdle" model was originally proposed by Mullaby (1986), i.e. $(1-\pi_i)$ is the probability of "clearing the hurdle" and generating a non-zero count. Both the binomial model and the count model in equation (9) can be changed. Hilbe (2008, Chapter 8) describes nine commonly used hurdle models. Camron and Trivedi (1998, Chapter 4) provide a good overview of hurdle models, their relationships to other count data models, and applications in economic studies.

If, however, it is assumed that zero counts can occur as the result of a "dual state process" then ZI models have been proposed to explain the excess zeros. Returning to the root propagation example above, the ZIP model is based on the assumption that there are two types of shoots. The first type of shoot gives a Poisson-distributed count which may be zero. The second type of shoot is not viable and always has zero roots (an assumption that is based on some biological principle for this type of plant cutting). The key issue is a logical justification for the assumption that structural zeros are possible, i.e. there is some set of unobservable conditions that occur for certain observational units that will always lead to a zero count. The resulting model is a discrete mixture distribution with a proportion ω of the counts being zero and the remainder of the counts following the Poisson distribution, i.e.

$$\Pr[Y_i = y_i] = \begin{cases} \omega_i + (1-\omega_i)\exp(-\lambda_i), & y_i = 0 \\ (1-\omega_i)\exp(-\lambda_i)\lambda_i \,\hat{}\, y_i)\,/y_i!\,, & y_i > 0\,, \end{cases} \qquad (10)$$

where the parameters λ_i and ω_i depend on vectors of covariates x_i and z_i, respectively. Equation (10) is the ZIP regression model originally proposed by Lambert (1992). ZI forms of the NB distribution (ZINB) are described by Ridout et al (2001) and Hilbe (2008, Chapter 8). Lord et al (2005) provide a detailed account of the application of Poisson, NB, ZIP, and ZINB regression models to the analysis of motor vehicle crash-data. They note that the ZI regression models (that have been used when the data contain too many zeros) assume that a dual-state process is responsible for generating the crash data. This model assumes that event counts are being generated from two different states, a "virtually safe" true zero-count state in which events never occur and a normal count-process state (which may happen to record zero events) that follow a Poisson or NB distribution. They use empirical data, theoretical principles, and simulation experiments to show that motor vehicle crash data characterized by a large number of zeros is not caused by a dual-state process and describe conditions that may explain the apparent excess zeros -- e.g., the omission of critical variables or the use of analysis sites characterized by a combination of low exposure, high risk, and heterogeneity. In a subsequent report (Lord, et al. 2007) they propose that the justification of ZI models based on improved statistical fit is not appropriate ("the maximizing statistical fit fallacy") unless the analyst can describe the conditions that characterize the two states. The logical problem (assuming that counted events are derived from a dual state process) is to describe unobserved conditions that could justify the two-state assumption. In highway safety studies (where the response variable is the number of crashes) this requires an explanation of observational units (e.g. a highway segment with known characteristics during a specified time period) that are in an inherently safe or non-safe state. Those observational units that have at least one event would be classified as unsafe; and the observational units with zero counts could be in either an inherently safe or unsafe state. If a logical basis for an "inherently safe" (i.e., a structural zero) state cannot be provided, then the dual state process is an artificial construct that has been used to improve the statistical fit. A ZI model does not seem reasonable unless it has been proposed a prior for the count variable of interest, e.g. the analyst should propose the ZIP model (10) for y_i and covariates of interest before data analysis.

Zeileis et al (2008) provide a detailed review of the conceptual and computational features of both the hurdle and ZI regression models. The R package **psc1** (Jackson, 2008) has functions **hurdle**( ) and **zeroinfl**( ) that can be used for data analysis. The design of these functions and the methods for the fitted model objects follows that of base R functions **glm**( ) in the **stats** package **glim.nb**( ) in the **MASS** package.

```
> #              APPENDIX 2. COMPUTATIONAL DETAILS
> #
> # POISSON REGRESSION ANALYSIS OF ILLNESS AND INJURY SURVEILLANCE DATA
> # Technical Report ORISE-09-OEWH-0176 (2012)   short name (PRA_IISP)
> #    Available at URL  http://www.csm.ornl.gov/~frome/priisp
> #    or mirror site http://home.comcast.net/~fromestat/priisp
> #               File Name readdata.txt at Step 7 (this file)
> #               To obtain results at Step 7 of website
> # Line 9        RUN R and open a txt file using sink(), i.e.
> #               sink("PRA_IISPdemo.txt")
> #               then  SOURCE THIS FILE WITH echo=TRUE, i.e.
> # line 12   source("readdata.txt",echo=TRUE,max.deparse.length =3000)
> #
> #  Input data frame drtw with IISP RTW data from binary file drtw.rda
>           load("./RFIISP/drtw.rda")

> #    with column names
> #    "yr","fa","sx","og","ag","pyr","all","rsp","inj","hrt","msc")
> #
> #  see technical report PRA_IISP in Item 1 for additional Information
>
> #     "yr","fa","sx","og","ag"       are factor covariates
> #     "all","rsp","inj","hrt","msc"  are counts of RTW health events
> #
> ####################################################################
> #
> #  The folder RFIISP contains utility functions that were used to
> #  obtain some of the numerical results in PRA_IISP--- see Example 1
> #  NOTE:
> #  THESE FUNCTIONS ARE PROVIDED FOR THE READERS CONVENIENCE AND
> #  ARE SPECIFIC TO THE IISP DATA STRUCTURES AND METHODS
> #-------------------------------------------------------------------
>   source("./RFIISP/read.RF.R")

>   ogn7<-read.RF()# input R functions and Occupational Group long names

>   library(MASS)   # attach MASS libraay

> ####################################################################
> #
> #     calculate marginal table of rsp events, person years,
> #     and rsp event rates for Occupational and  Age Groups
> #     for SNL--- see Table 1 in PRA_IISP report
> #--------------------------------------------------------------------
>    evcount.oa(d=drtw,rfa="SNL",ryr="ALL",event="rsp",pc=F,ognames=ogn7)

Number of  RSP RTW Absences  By Occupation and Age For SNL Year= ALL

                       AGE 16-29 30-39 40-49 50+
OCCUPATION
A-Administrative             11    72   156   171
C-Crafts                      2    17    34    31
E-Fire and Security           1     7    11     3
```

```
O-Line Operators               1     0     5     4
P-Professionals(F I M)         7    78   184   163
S-Service                      2    20    35    30
T-Technical Support-B          5    37    78    61
------------------------------------------------------------

Person Years By Occupation and Age
                       AGE 16-29 30-39 40-49 50+
OCCUPATION
A-Administrative            1697  3801  6446  5907
C-Crafts                      95   467  1215  1061
E-Fire and Security           96   466   445   201
O-Line Operators              66   315   398   186
P-Professionals(F I M)      1980  9250 14455 13030
S-Service                    210   342   771   758
T-Technical Support-B        583  2330  4124  2986
------------------------------------------------------------

 RSP  Event Rate Per 1000 Person Years
                       AGE 16-29 30-39 40-49 50+
OCCUPATION
A-Administrative             6.5  18.9  24.2  28.9
C-Crafts                    21.1  36.4  28.0  29.2
E-Fire and Security         10.4  15.0  24.7  14.9
O-Line Operators            15.2   0.0  12.6  21.5
P-Professionals(F I M)       3.5   8.4  12.7  12.5
S-Service                    9.5  58.5  45.4  39.6
T-Technical Support-B        8.6  15.9  18.9  20.4
------------------------------------------------------------

> ######################################################################
> #
> #     Fit the Main Effects Model(MEM) to RTW rsp event data
> #     --- see Table 2 in PRA_IISP Technical Report
> #---------------------------------------------------------------------
>   h <- make.h(drtw,rfa="SNL",event="rsp")

>   make.h  #  this function extracts an event at facility rfa
function(d=drtw,rfa="SNL",event="rsp"){
#      make data frame for Poisson regression using
#      constraints described in PRA_IISP report
#      Technical Report ORISE-09-OEWH-0176 (2012)
#
        if(rfa!="ALL")d<-d[d$fa==rfa,]
           y <- d[[event]]
           nn<- d$pyr/1000
      h <-  data.frame(y,nn,d[,1:5])
     contrasts(h$ag)<-contr.treatment(4,3)
     contrasts(h$sx)<-contr.treatment(2,2)
     contrasts(h$og)<-contr.sum(7)
     contrasts(h$yr)<-contr.sum(10)
h
}
```

```
> ######################################################################
> #
> #  fit the MEM using R function glm()--- see glm help file for details
> #
> #--------------------------------------------------------------------
>      mf<- glm(y ~ ag+sx+og+yr,fa=poisson,data=h,offset= log(nn) )

>      summary(mf)

Call:
glm(formula = y ~ ag + sx + og + yr, family = poisson, data = h,
    offset = log(nn))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7576  -0.6823  -0.3376   0.4012   2.9285

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.84978    0.07284  39.126  < 2e-16
ag1         -1.24668    0.19173  -6.502 7.92e-11
ag2         -0.30478    0.08009  -3.805 0.000142
ag4          0.07349    0.06459   1.138 0.255229
sx1          0.72337    0.06447  11.220  < 2e-16
og1         -0.14766    0.07907  -1.868 0.061829
og2          0.50737    0.11151   4.550 5.36e-06
og3          0.13896    0.19095   0.728 0.466771
og4         -0.53522    0.27418  -1.952 0.050931
og5         -0.60470    0.07300  -8.284  < 2e-16
og6          0.70041    0.10933   6.407 1.49e-10
yr1          0.40790    0.07235   5.638 1.72e-08
yr2          0.31967    0.07569   4.223 2.41e-05
yr3         -0.03362    0.09136  -0.368 0.712891
yr4         -0.20668    0.09590  -2.155 0.031148
yr5          0.17161    0.08159   2.103 0.035429
yr6         -0.13629    0.09330  -1.461 0.144088
yr7         -0.10801    0.09209  -1.173 0.240847
yr8         -0.11033    0.09057  -1.218 0.223142
yr9         -0.09401    0.08880  -1.059 0.289787

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 937.16  on 536  degrees of freedom
Residual deviance: 485.43  on 517  degrees of freedom
AIC: 1380.4

Number of Fisher Scoring iterations: 6


> #####################################################################
> #
> #  calculate analysis of deviance table for facility SNL rsp events
> #            see Table 3 in PRA_IISP report
> #---------------------------------------------------------------------
>      aodme4(drtw,"SNL")
       q Deviance     IC Df     LRT  LogO
MEM   20   485.43 565.43 NA      NA    NA
ag    17   570.07 638.07  3   84.64 40.31
og    14   656.74 712.74  6  171.31 77.42
yr    11   545.25 589.25  9   59.81 20.35
sx    19   608.40 684.40  1  122.96 64.12
ag:og 38   462.04 614.04 18   23.39  1.54
ag:yr 47   453.21 641.21 27   32.23  1.24
ag:sx 23   480.56 572.56  3    4.87  1.51
og:yr 74   425.92 721.92 54   59.51  0.93
og:sx 26   473.13 577.13  6   12.31  2.84
yr:sx 29   475.67 591.67  9    9.77  0.53

> #####################################################################
> #
> #      Use utility function compare.estimates() --- see listing below
> #      TO COMPARE PARAMETER ESTIMATES AND STANDARD ERRORS
> #      FOR FOUR METHODS BASED ON IWLS ALGORITHM
> #
> #      1) Poisson ML    2) QL/M1     3) NB ML and    4) QL/M2
> #
> #      for IISP data for facility= "PTX"  event= "rsp"
> #
> #      THIS IS EXAMPLE 2 IN TECHNICAL REPORT PRA_IISP.
> #-------------------------------------------------------------------
>     compare.estimates(drtw,rfa="PTX",event="rsp",mf=NA)
$coef
             ML-Po  QLM_1   ML_NB  QLM_2
(Intercept)  3.282  3.282   3.310  3.315
ag1         -0.404 -0.404  -0.414 -0.416
ag2          0.186  0.186   0.172  0.166
ag4          0.180  0.180   0.158  0.156
sx1          0.545  0.545   0.515  0.510
og1         -0.262 -0.262  -0.321 -0.337
og2          0.159  0.159   0.188  0.196
og3         -0.124 -0.124  -0.112 -0.105
og4          0.605  0.605   0.604  0.604
og5         -1.102 -1.102  -1.127 -1.134
og6          0.407  0.407   0.448  0.457
yr1         -0.207 -0.207  -0.210 -0.212
yr2          0.011  0.011   0.010  0.007
yr3         -0.387 -0.387  -0.394 -0.397
yr4         -1.433 -1.433  -1.422 -1.417
yr5         -0.852 -0.852  -0.844 -0.841
yr6         -0.974 -0.974  -0.970 -0.969
yr7          0.264  0.264   0.259  0.256
yr8          0.924  0.924   0.908  0.902
yr9          1.384  1.384   1.388  1.392

$ster
            ML-Po QLM_1   ML_NB QLM_2
(Intercept) 0.059 0.066   0.068 0.072
ag1         0.114 0.128   0.127 0.132
```

```
ag2         0.069 0.077   0.083 0.089
ag4         0.063 0.070   0.079 0.085
sx1         0.059 0.066   0.070 0.074
og1         0.058 0.065   0.071 0.076
og2         0.080 0.089   0.093 0.099
og3         0.065 0.073   0.079 0.084
og4         0.051 0.058   0.065 0.070
og5         0.076 0.085   0.088 0.093
og6         0.086 0.097   0.101 0.107
yr1         0.104 0.116   0.114 0.118
yr2         0.094 0.105   0.105 0.109
yr3         0.112 0.126   0.122 0.126
yr4         0.195 0.218   0.200 0.202
yr5         0.148 0.167   0.156 0.159
yr6         0.158 0.178   0.165 0.168
yr7         0.088 0.099   0.099 0.104
yr8         0.068 0.076   0.082 0.087
yr9         0.060 0.067   0.074 0.080

$disp
[1] 1.00000 0.25876 0.07183 0.10762

$m2ll
        ML-Po QLM_1   ML_NB QLM_2
logLik 1597.8    NA  1578.9    NA
Df       20.0    20    21.0    20

$mfit
[1] "rsp events at PTX y ~ ag + sx + og + yr"


> #####################################################################
> #
> #      List R utility function compare.estimates()
> #-------------------------------------------------------------------
>    compare.estimates
function(d=drtw,rfa="PTX",event="rsp",mf=NA){
#
#     COMPARE PARAMETER ESTIMATES AND STANDARD ERRORS
#     for four methods based on IWLS algorithm
#
#       1)Poisson ML 2) QL/M1 3) NB ML and 4) QL/M2
#  for IISP data for facility= rfa heath event=event model = mf
#     [see PRA_IISP equation (6) and related text for details ]
#
# ARGUMENTS:
#           d data frame with IISP data
#           rfa facility
#           event health event for summary table
#           mf model DEFAULTS to y~ ag +sx+og+yr
#
# USAGE: compare.estimates(drtw,rfa="PTX",event="rsp",mf=NA)
#
# VALUE: a list with components
#       coef  parameter estimates for each method
#       ster  estimated standard errors for each method
#       disp  estimate dispersion parameter
#       m2ll -2*loglikelihood (for models with likelihood)
#       mfit  describes data and model fit
#
#  REQUIRES utility functions
#        make.h(d,rfa,event)
#        iwls.qm2(mf,h)
#
#        make.h() to make data frame for regression
      h<- make.h(d,rfa,event)
#        determine model formula ( see glm help file)
      if(is.na(mf) ) mf<-"y ~ ag +sx+og+yr"
      fmp <- as.formula(mf)
      fmnb<- as.formula( paste(mf,"+offset(log(nn))") )

#        fit Poisson model
   f_p   <- glm(fmp , fa=poisson, data = h,offset= log(nn) )
#        fit QL/M1 quasi-likelihood method of moments 1 model
   f_qm1 <- glm(fmp, fa=quasipoisson, data = h,offset= log(nn) )
#        chi-sq dispersion statistic for QL/M1
   phi.p<- sum( residuals(f_qm1,type="pearson")^2 )/f_qm1$df.residual
#        fit NB model using glm.nb in package MASS
   f_nb<-glm.nb( fmnb,data=h,maxit=100,epsilon=1e-05)
#        fit QL/M2 quasi-likelihood method of moments 2 model
#        see comments in utility function iwls.qm2()
   f_a2 <-iwls.qm2(mf,h)
   f_qm2 <- f_a2$f
#
    fm <- list("ML-Po" =f_p,"QLM_1"=f_qm1  ,"  ML_NB" = f_nb,"QLM_2"=f_qm2
)
    np<-length(f_p$coef)
#########  coefficients for each model
coef<-  round( sapply(fm, function(x) coef(x)[1:np]), 3)

#########   standard errors for each model
    ster <- round( sapply(fm, function(x) sqrt(diag(vcov(x)))[1:np]),3)
    ster <- round(ster,3)
#########   calculate dispersion parameter for each model
disp<-round( c(1,phi.p-1,1/f_nb$theta, f_a2$delta),5 )
#########  calculate  log-likelihood for Poisson and NB models
logl<-rbind(logLik = -2*sapply(fm, function(x) round(logLik(x), digits =
2)),
  Df = sapply(fm, function(x) attr(logLik(x), "df")))
    logl[1,4]<- NA
    mfit<-paste(event,"events at",rfa,deparse(f_p$formula) )
out<-list(coef=coef,ster=ster,disp=disp,m2ll=logl,mfit=mfit)
out
}
> #
> #    Close file PRA_IISPdemo.txt and exit R Dec 1 2012
> #----------------------------------------------------------------------
>      sink()
```