COENOSES 10(2-3):113-121, 1995.

ASSESSMENT OF AQUATIC SYSTEMS TOXICITY USING GENERALIZED LINEAR MODELS AND QUASI-LIKELIHOOD TECHNIQUES[1]

Amarjot Kaur, Dario Gregori, G. P. Patil and C. Taillie

Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA

Keywords: Generalized Estimating Equations, Generalized Linear Models, Quasi-Likelihood, Toxicity Assessment, Water Pollution.

Abstract. Assessing pollution impact on water ecosystems is of great concern. Modelling such impact is usually done in a regression framework on a suitable transformed variable, under somewhat strict assumptions on residual distribution and variability. Main aim of this paper is to illustrate some alternative approaches in the framework of Generalized Linear Models. In particular, distributional-free, semiparametric approaches, such as quasi-likelihood and generalized estimating equations, will be discussed with reference to an application on reproductive output in fresh water zooplankton.

1. Introduction

Protection of the aquatic environment has been a matter of great concern and a number of legislative acts aiming to control water pollution have been enacted. Pollutants from several sources, such as fertilizers, synthetic pesticides, and other industrial and municipal effluents may disrupt the aquatic life of various vertebrates, invertebrates and plants. Generally, they adversely affect growth and reproduction as well as leading to an increased incidence of birth defects. Several studies have been conducted to assess the lethal and behavioral impact of toxicants, and interestingly, different data analysis techniques can be applied to the same dataset. In this paper, we discuss some of the methods available from the realm of generalized linear regression and quasi-likelihood, which are subsequently illustrated using a dataset of reproductive output in freshwater zooplankton.

Traditional techniques of linear regression and analysis of variance are somewhat limiting due to the requirements of constant variance and distributional assumptions. Generalized linear models (GLM) extend the range of application by relaxing the requirements of classical linear models. For an overview of the applications of GLM in ecology, see Crawley 1993) and Kaur et al. (1995).

Linear least squares regression ranks among the most versatile and often used methods of data analysis in applied statistics. However, some of classical assumptions can limit its applicability. For example, two major limitations are (i) the mean response is a linear function of the regression parameters and (ii) the error variance is the same for all observations. Normality is needed, but only to establish the exact, small sample distribution theory.

The shortcomings of classical regression are addressed by GLM in two ways. First, the mean of the response variable is transformed rather than the response variable itself. The transformed mean is then modeled as a linear combination of the covariates with unknown regression coefficients. Second, the variance of the response variable is taken to be proportional to some known function of its mean. Depending upon the form of the variance function there may exist a linear exponential family that is compatible with the moment model. Then, maximum likelihood could be used for further inference. However, the needed linear exponential family does not always exist, especially when the dispersion parameter is needed to account for *overdispersion* in the data.

Wedderburn (1974) developed a method of inference, known as *quasi-likelihood*, that does not require the specification of a distribution for the response variable. If the moment model happens to have a corresponding linear exponential family and if the data come from that family, then the quasi-likelihood estimates are the same as the maximum likelihood estimates. In this distribution-free respect, quasi-likelihood is similar to least squares where the normal distribution plays the role of the linear exponential family. In a more general situation, *extended quasi-likelihood* and

1 Prepared with partial support from the United States Environmental Protection Agency, Environmental Monitoring and Assessment Program, EMAP Design and Statistics Group under a Cooperative Agreement Number CR-821783. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

pseudo-likelihood approaches allow dispersion parameters to vary from observation to observation. Extending the scope of quasi-likelihood estimating equations further, Liang and Zeger (1986), proposed the *generalized estimating equations* (GEE), as a way to approach data in the form of correlated repeated measures.

Some techniques of GLM and their background are outlined in Section 2, and their illustration is made through an example, using a real data set (Bailer and Oris 1994) in Section 3.

2. Background and Methods

This section can be broadly classified into three different categories: generalized linear models, quasi-likelihood techniques, and generalized estimating equations. Let Y be the response variable whose mean,

$$\mu = E[Y] \tag{1}$$

varies according to the values of some available covariates $x_1, x_2, ..., x_p$. The dependence of μ upon the covariates is modeled by the equation

$$g(\mu) \equiv \eta = \beta_1 x_1 + ... + \beta_p x_p = \beta' \mathbf{x}, \tag{2}$$

where $g(\cdot)$ is a known transformation called the *link function* and where $\beta_1, ..., \beta_p$ are unknown regression coefficients. The quantity $g(\mu)$ is referred as the linear predictor. The variance of Y is taken to be proportional to some known variance function $V(\cdot)$ of μ. Thus,

$$\text{var}(Y) = \varphi V(\mu), \tag{3}$$

where φ is called the dispersion parameter which could be a known or an unknown constant.

The variance to mean ratio describes the variability in the data and is used to characterize different families of distribution. The GLM framework allows the investigator to specify particular variance to mean relationships and to draw inferences accordingly. Familiar examples include:

- Binomial proportions with $V(\mu) = \mu(1-\mu)$ and $\varphi = 1/m$.
- Poisson counts with $V(\mu) = \mu$ and $\varphi = 1$.
- Constant coefficient of variation (CV) with $V(\mu) = \mu^2$ and $\varphi = CV^2$.

In the first two examples, the specific values for φ can be replaced by unknown φ to represent underdispersion $\varphi<1$ or overdispersion $\varphi>1$ in the response variable.

Binomial Variance to Mean Relation

Logistic regression and Ordinal Response Models

The logistic regression is suitable for distributions having binomial variance to mean ratio. The logit link between expectation and linear predictor is given by $\text{logit } \pi \equiv \log\left(\frac{\pi}{1-\pi}\right) = \beta' \mathbf{x}$ Much of the popularity of the logit link is due to the fact that it represents the log-odds as an additive function of the covariates so that the odds itself depends multiplicatively on the covariates. Logistic regression has been widely used in ecological work and enjoys myriad applications. See McDonald et al. (1992), Futter (1994), and Kaur et al. (1995).

Another approach can be followed by modeling the cumulative distribution of the response variable. This approach is meaningful for ordinal categorical responses and takes the ordering of the response variable into account. Let Y be an ordinal response variable having K categories. Let $F_k(\mathbf{x}) = Pr(Y \geq k|\mathbf{x})$ be the cumulative probabilities greater than equal to category k, conditional upon the explanatory variables. We can write a model

$$\Pr\{Y \geq k|\mathbf{x}\} = \frac{1}{1 + \exp[-(\alpha_k + \beta' \mathbf{x})]}, \quad k = 1, 2, ..., K, \tag{4}$$

or, equivalently,

$$L_k(\mathbf{x}) \equiv \text{logit}(F_k(\mathbf{x})) = \alpha_k + \beta' \mathbf{x}$$

where the α_k are *cutpoint* parameters, nondecreasing in k for fixed $\mathbf{x}$. In general, the log of the cumulative ($\geq k$) odds ratio of two different values of the explanatory variables is given by

$$L_k(\mathbf{x}_1) - L_k(\mathbf{x}_2) = \log \frac{P(Y \geq k|\mathbf{x}_1)/P(Y < k|\mathbf{x}_1)}{P(Y \geq k|\mathbf{x}_2)/P(Y < k|\mathbf{x}_2)}, \tag{5}$$

For model (4), this simplifies to

$$L_k(\mathbf{x}_1) - L_k(\mathbf{x}_2) = \beta'(\mathbf{x}_1 - \mathbf{x}_2),$$

which is same for all cumulative categories k. In other words, the odds of making response $\geq k$ are $\exp(\beta'(\mathbf{x}_1 - \mathbf{x}_2))$ times higher at $\mathbf{x} = \mathbf{x}_1$ than at $\mathbf{x} = \mathbf{x}_2$.

Poisson Variance to Mean Relation

Log-linear models

Log-linear models that are often applied to the analysis of count data, typically using a Poisson variance to mean ratio, see Bishop and Holland (1975). A primary feature of the log-linear relationship, $\log \mu = \eta = \beta' \mathbf{x}$, is to represent the dependence of the expectation $\mu = E(Y)$ of the counting process upon the covariates. The log-link insures the positivity of all the fitted values. The log-linear representation has a strong connection with the theory of contingency tables Roy and Kastenbaum (1956). This idea apparently arose from the work of Yule (1912) who was a proponent of the use of *odds ratios* as the building blocks of association and interaction in contingency tables. Population size estimation is one of the major application of log-linear models. Schwartz (1991, 1994), and Pope et al. (1992) discuss various applications of log-linear regression for assessing the geographical variation of pollutants across population aggregates such as, farms, cities, etc. This approach allows a very flexible control for time trends, seasonal fluctuations, and weather in assessing the behavior of the response, as depending on the number of inhalable particles (in air pollution studies) or concentration levels. Sections 3.1 and 3.2 describe further detailed application of log-linear models.

Quasi-Likelihood

This method of estimation and inference was first introduced by Weddurburn (1974), and subsequently developed

by McCullagh (1983). A comprehensive review with applications may be found in McCullagh and Nelder (1989). Let the variance of the response variable Y be proportional to some known *variance function* $V(\cdot)$ of the mean μ. For a single observation y, the quasi-likelihood $Q(\mu;y)$ is defined to be any solution of the differential equation

$$\frac{\partial Q}{\partial \mu} = \frac{y-\mu}{\varphi V(\mu)}. \tag{6}$$

The total quasi-likelihood (TQL) is obtained by summing over all observations,

$$\sum_{i=1}^{n} Q(\mu_i;y_i). \tag{7}$$

Because of equation (2), the TQL is a function of the regression coefficients $\beta_1, ..., \beta_p$. The maximum quasi likelihood (MQL) estimates are the values of $\beta_1, ..., \beta_p$ that maximize TQL. These are obtained as the solution of the equations

$$\sum_{i=1}^{n} \frac{y_i-\mu_i}{\varphi V(\mu_i) g'(\mu_i)} x_{ir} = 0, \quad r = 1, 2, ..., p, \tag{8}$$

where $x_{ir} = \partial\eta_i/\partial\beta_r$ is the value of the rth covariate for the ith observation. If the dispersion parameter φ is unknown, it is generally estimated as

$$\hat{\varphi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i-\mu_i)}{V(\hat{\mu}_i)}. \tag{9}$$

In Wedderburn's formulation, the applicability of quasi-likelihood is subject to two restrictions: (i) The dispersion parameters φ_i need to be the same for all observations. More generally, $\varphi_i = \varphi a_i$ where a_i is known. (ii) The responses $Y_1, Y_2, ..., Y_n$ are independent. These restrictions can be addressed by extended quasi-likelihood and by generalized estimating equations, respectively.

Overdispersion Diagnostics

The presence of greater variation than expected for a nominal model is known as *overdispersion*. It is important to allow for *overdispersion* in the model in order to obtain correct variance estimates and valid hypothesis tests. GLM methods account for *overdispersion* by introducing an additional parameter into the estimating equations. Let Y_i, $i = 1, ..., n$, be independent response variables satisfying $E(Y_i) = \mu_i$ with $g(\mu_i) = \eta_i = \beta'\mathbf{x}_i$ and $\text{var}(Y_i) = \varphi_i V(\mu_i)$, where $V(\cdot)$ is a known function and φ_i is a dispersion parameter. In the quasi-likelihood setting, the φ_i are known up to a single unknown multiplicative factor and are assumed to vary proportional to known weights a_i. Extended quasi-likelihood allows for multiple unknown dispersion parameters, which can be modeled as

$$\text{var}(Y_i) = \varphi_i a_i V(\mu_i) \text{ with } h(\varphi_i) = \lambda + \alpha'\mathbf{z}_i$$

where the a_i are known constants, λ is a scalar parameter, and $\mathbf{z}_i$ and α are $q\times 1$ vectors of explanatory variables and of unknown parameters, respectively. If $\alpha = \mathbf{0}$ then φ_i is a constant dispersion parameter.

Many tests have been proposed for detecting overdispersion and for modeling any extra-variation detected in the data. For details, see Efron (1986), Nelder and Pregibon (1987) and Carroll and Ruppert (1988). It is often convenient to represent the overdispersion in terms of the reciprocal of φ, the reciprocal being an 'information' parameter. Ganio and Schafer (1992) discuss the effect of carcinogenic toxicants on rainbow trout. The extra-binomial dispersion indicated in the data is assessed using different models for the overdispersion: (i) constant, (ii) separate dispersion parameter for each toxicant and (iii) random effects model.

We observe in Section 3 that different models result in different values of overdispersion, explaining the unaccountable variation of that particular model.

GEE

Liang and Zeger (1986) proposed the *generalized estimating equations* (GEE) as a way to approach data in the form of correlated repeated measures (e.g., longitudinal data) in a semi-parametric framework, which has received considerable attention. See Zeger and Liang (1992) for an overview of GEE. In a longitudinal data framework, the subjects are independent because of the design, but generating correlated observations. The basic idea in GEE is the generalization of the quasi-likelihood estimating equations to allow a block-diagonal covariance matrix of the response vector.

Let the response variable Y_{ij} correspond to the jth observation on the ith individual, having corresponding explanatory variables $\mathbf{x}_{ij}$ (vector of size p); $j=1, ..., n_i$; $i = 1, .., K$. A marginal GEE model is based on the same equations as the GLM, namely $E(Y_{ij}) = \mu_{ij}$ and $\text{var}(Y_{ij}) = \varphi V(\mu_{ij})$. Also $g(\mu_{ij}) = \eta_{ij} = \beta'\mathbf{x}_{ij}$. The covariance between Y_{is} and Y_{ih} is modeled as $\text{cov}(Y_{is},Y_{ih}) = c(\mu_{is},\mu_{ih};\alpha)$, where $1 \le s \le h \le n_i$, for some α. Also, we take $g(\mu_{ij}) = \eta_{ij} = \beta'\mathbf{x}_{ij}$.

Under mild regularity conditions, Liang and Zeger (1986) prove that the GEE estimator for β is given by

$$U_\beta(\hat{\beta}) = \sum_{i=1}^{K} D_i^T V_i^{-1} S_i = 0, \tag{10}$$

where $S_i = Y_i - \mu_i$, $\mu_i = (\mu_{i1}, ..., \mu_{in_i})^T$, $D_i = (\partial\mu_i/\partial\beta)$ and $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$, where A_i is an $n_i \times n_i$ diagonal matrix with $a_{ii} = \text{var}(Y_{ij})$, $j = 1, ..., n_i$ and $R_i(\alpha)$ a working correlation matrix. Both $\hat{\beta}$ and the estimator of its variance are consistent even if V_i is not correctly specified.

One of the main advantages of the GEE method is that it allows the time dependence to be specified in a variety of ways. Some common specifications for the correlation matrix $\rho(Y_i)$ are as follows:

- $R_i(\alpha) = \mathbf{I}$, where $\mathbf{I}$ is an $n_i \times n_i$ identity matrix. This corresponds to the *working independence* assumption, which yields to estimating equations identical to the ordinary logistic regression equations.
- Exchangeable correlation: $\rho(Y_{is},Y_{it}) = \alpha$ for each $s \neq t$.

- Autoregressive correlation: $\rho(Y_{is},Y_{it}) = \alpha^{|s-t|}$ for each $s \neq t$.
- Unstructured or, pairwise correlation: $rho(Y_{is},Y_{it}) = \alpha_{st}$, a member of $n_i(n_i-1)/2 \times 1$ vector containing all the pairwise correlations.

3. Toxicity Assessment in Aquatic Systems: An Example

In an effort to protect resources from industrial and municipal effluents, an increasing number of toxicity tests are conducted by the regulatory agencies. The effect of toxicants on growth and reproduction are the most common endpoints in these tests. A case study by Bailer and Oris (1994) dwells on reproductive toxicity response in *Cerodaphnia dubia* (a fresh-water zooplankton species) to nitrofen (a herbicide used to control grass weeds in cereal and rice). The test was conducted at five different concentration levels of nitrofen, where 10 m *Cerodaphnia dubia* each were exposed to different doses of nitrofen. The response variable Y is the observed number of offspring born in three broods to each of 10 *Cerodaphnia dubia* under different concentration levels. Let Y_{ijk} be the number of offspring corresponding to the ith (i=1, ..., I) dose level, jth (j = 1, ..., J) individuals and kth (k = 1, ..., K) brood. In this experiment $I = 5$, $J = 10$, $K = 3$. Bailer and Oris (1994) used log-linear model for this data, with response taken as the total number of offspring in the three broods. Their model does not take into account distinct behavior of brood 1 from brood 2 and brood 3 as seen from Figure 1. In this section, we discuss various alternative approaches of analyzing the same data, which could be more meaningful in light of the special features of the data.

Aggregated Response Model (Bailer and Oris 1994)

As response variable, Bailer and Oris (1994) used the total number of offspring per subject, obtained by summing them over the three broods. Thus,

$Y_{ij} = Y_{ij1} + Y_{ij2} + Y_{ij3}$,

with

$E(Y_{ij}) = \mu_i$

and

$V(Y_{ij}) = \varphi\mu_i$.

With C_i as the concentration level, the expectation μ_i is modeled as an exponential function of an M-degree polinomial

$$\mu_i = \exp\left(\sum_{m=0}^{M} \beta_m C_i^m\right). \quad (11)$$

An appropriate value of M is chosen by making an evaluation of the residuals. The model proposed by Bailer and Oris (1994) used a quadratic polynomial of the dose (Table 1), which leads to an estimated RI50 of 240 µg/L (RI50 is the dose level at which 50% of the population is affected). In order to assess the non-linearities not accounted for by the quadratic model, an alternative has been considered in which the dose is treated as a five-level factor variable (Table 2). In other words, each of the five dose levels is allowed to have its own separate mean response.

Table 1. Bailer and Oris (1994) model for the aggregate response, with dose as continuous covariate. The residual deviance 55.9 on 47 d.f. and $\hat{\varphi}$ = 1.09.

Coefficient	Value	Std. Error	*t* value
(Intercept)	3.409	0.054	62.165
Dose	0.0034	0.001	3.937
Dose^2	0.000	0.000	8.193

Table 2. Factor model for the aggregate response, with dose as categorical covariate. The residual deviance 50.71 on 45 d.f. and $\hat{\varphi}$ = 1.05

Coefficient	Value	Std. Error	*t* value
(Intercept)	2.975	0.0369	80.529
Dose (level1)	0.001	0.040	0.038
Dose (level2)	-0.035	0.024	-1.436
Dose (level3)	-0.142	0.021	-6.657
Dose (level4)	-0.295	0.027	-10.846

Table 3. Analysis of deviance for the quadratic and factor models. Terms added sequentially.

Model	Df	Deviance	Resid. Df	Resid. Dev
Quadratic				
NULL			49	312.48
Dose	1	184.89	48	127.59
Dose^2	1	71.75	47	55.87
Factor				
NULL			49	312.48
Dose	4	261.77	45	50.72

To select between the two models, a likelihood ratio test can be conducted based on the change in deviance between the models. Table 3 shows the analysis of deviance table for both models. The first approach may be preferred for its ease of interpretation, whereas the latter may be preferred from the modeling point of view.

The estimated overdispersion parameters $\hat{\varphi}$ obtained in the two modes are 1.09 and 1.05 respectively. Thus, there is no indication of extra-binomial dispersion in the total number of offspring per subject. Some further issues may be raised after having a closer look at both the data and the model specification of Bailer and Oris (1994). Figure 1 shows that the behavior of the responses is very different for different brood levels, collapsing data into one response can be very misleading from the point of view of modeling and,

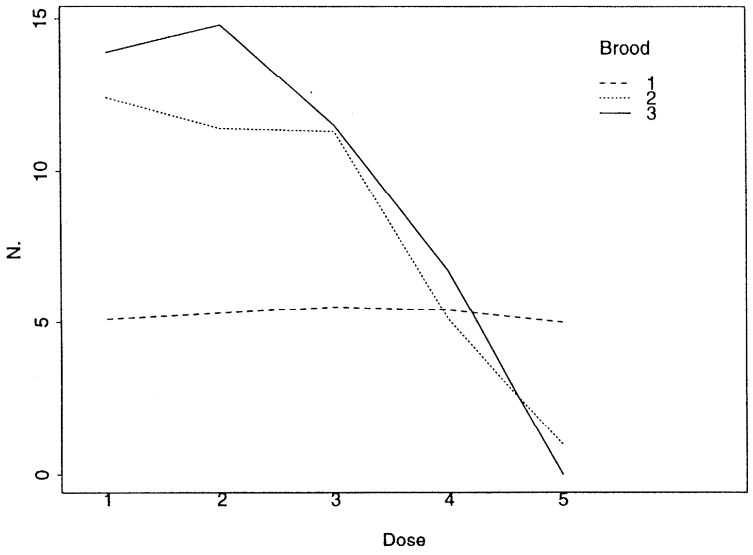

Figure 1. ***Cerodaphnia dubia*** **data. Interaction plot (least square fitted lines on three broods separately) of the response to evaluate the brood effect;** ***x-axis:*** **dose, and** ***y-axis:*** **number of offspring.**

more generally, leads to loss in information. In particular, as can be seen from Figure 1, brood one has a very different response from broods two and three, which are very similar. The different response behavior for the three broods, and in particular for the first as compared with the last two, can be modeled in several ways.

Disaggregated Response Models

One of the most straightforward approches is to consider the same model as before, but for the raw vector of the data, in which the broods are lined up. In this case, the mother effect becomes important, inducing a clustering effect not milded by collapsing (unlike Bailer's model), and thus inflating the variance by an overdispersion parameter. The correlation induced by the clustering of offsprings can be reparametrized, following McCullagh and Nelder (1989) as a function of the overdispersion parameter φ.

The expectation of the *raw* response Y_{il} can be modeled as

$$\mu_i = \exp\left(\sum_{m=0}^{M} \beta_m C_i^m\right), \qquad (12)$$

where $l = 1, ..., (K \times J)$. In fact the dispersion parameter can take into account the dependence in the data, without the loss in information due to collapsing.

The fitted model, as shown in Table 4, yields an estimate of the RI50 as 241 µg/L. This estimate is close to that obtained by Bailer and Oris (1994). A comparison of the two

Table 4. Model for disaggregated data without brood effect. Estimated overdispersion $\hat{\varphi} = 2.18$.

Coefficients	Value	Std. Error	*t* value
(Intercept)	2.311	.081	28.57
DOSE	0.004	.001	2.67
$DOSE^2$	0.000	0.000	-5.57

Table 5. Model for Brood 1 only data. Estimated underdispersion $\hat{\varphi} = 0.45$.

Coefficients:	Value	Std. Error	t value
(Intercept)	1.659	0.041	40.135
DOSE1	0.019	0.066	0.292
DOSE2	0.019	0.037	0.503
DOSE3	0.005	0.026	0.182
DOSE4	-0.013	0.021	0.595

Table 6. Model for Brood 2 and 3 combined data. Estimated overdispersion $\hat{\varphi} = 2.96$.

Coefficients	Value	Std. Error	t value
(Intercept)	1.796	0.131	13.680
DOSE1	-0.002	0.075	-0.025
DOSE2	-0.047	0.046	-1.032
DOSE3	-0.188	0.043	-4.414
DOSE4	-0.607	0.120	5.538

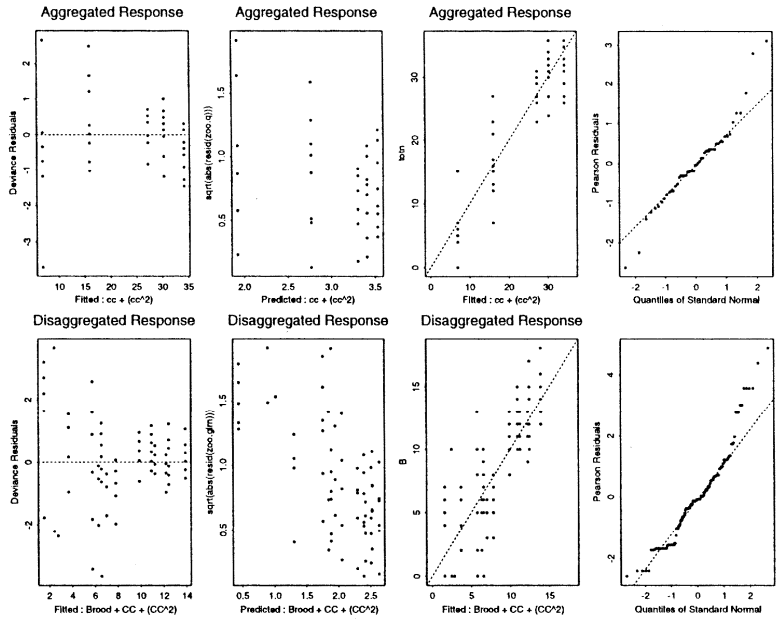

Figure 2. Aggregated model (Bailer and Oris 1994) compared with the disaggregated model (proposed).

models has been performed in Figure 2. The graph in the first column shows the fitted against the deviance components. As expected, the data in the new model show a greater variability, given essentially by disaggregation the data. The second column plots the linear predictor $\beta'\mathbf{x}$ against the absolute residuals $|Y - \hat{Y}|$, whereas the third column displays the fitted against the observed values in the two models.

Independent Brood Model. As shown in Figure 1, the brood one response is different from that of brood two and three. There is no indication of dose dependence for the first brood while brood size decreases sharply with increasing dose for the second and third broods. This suggests separate modeling of brood one response and of total response for brood two and three, combined. Tables 5 and 6 provide the fits for the factors model in the two situations. As expected, the two models fit very different. In addition, the fitted model for brood one shows a high degree of underdispersion ($\hat{\varphi} = .45$) while that of brood two and three shows high overdispersion ($\hat{\varphi} = 2.96$).

Mean Model. Recognizing the different behavior of brood one, only brood two and three were considered in the preceding section. These combined data were fitted using the model by Bailer and Oris (1994). It is seen from Table 6 and also Figure 3 that this fit is unsatisfactory in comparison with the aggregated model (Tables 1 and 2). Here, we consider another extension of their model, which incorporates the brood effect into the expectation, as

$$\mu_{ik} = \exp\left(\sum_{m=0}^{M} \beta_m C_i^m + \gamma B_k\right), \quad 1 = 1, 2, \ldots, 5; \; k = 1, 2, 3, \tag{13}$$

where B_k is a tricotomic variable indicating the brood. Assuming independence in the observation given the brood (i.e., assuming the mother's clustering effect to be negligible), we fitted the quasi-likelihood model given in Table 7 and Table 8, respectively, considering the dose as continuous and categorical variable.

Ordinal Response Models

The main idea behind Bailer's approach is that the initial dose effect can be fully evaluated only after a complete estimate of his cumulative (endpoint) effects.

A less conservative approach can be followed modeling the cumulative distribution of the response variable. This approach takes into account the natural ordering of the response variables. In this example, the three broods from each mother have underlying chronological ordering. The variable Y may be considered as an ordered sequence of 1, 2 or 3's, according to the brood they belong to, conditioning to the total number of offspring actually observed. The model is described in Section 2.1. The results of this model are given in Table 9.

Correlation Model

An alternative approach to the previous two is based on generalized estimating equations (GEE), in which the cor-

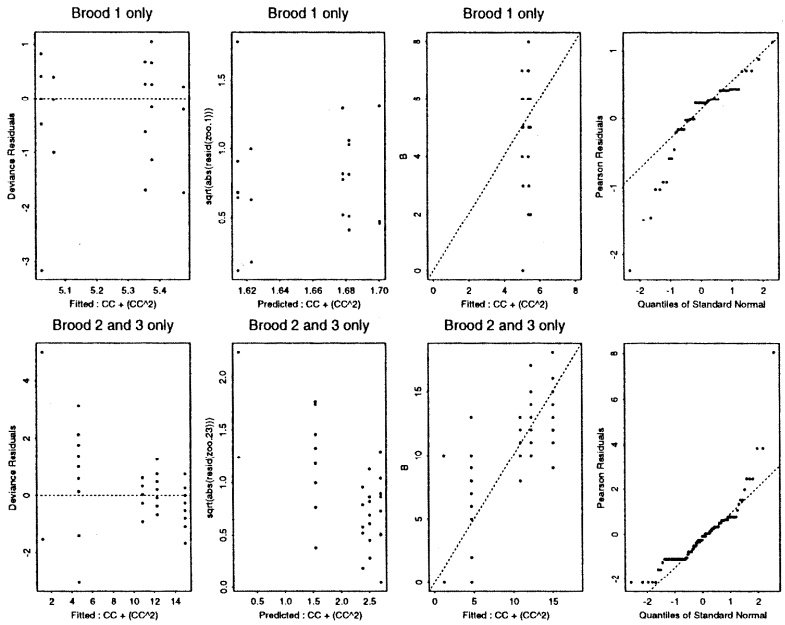

Figure 3. Diagnostics for the separate models. (i) Brood 1 only, and (ii) broods 2 and 3 together.

Table 7. Model for the mean as function of the brood. Estimated overdispersion $\hat{\varphi}$ = 1.97. Dose as continuous variable.

Coefficients	Value	Std. Error	*t* value
(Intercept)	1.850	0.191	9.677
Brood1	0.224	0.055	4.054
Brood2	0.118	0.028	4.155
DOSE	0.598	0.157	3.798
$Dose^2$	-0.162	0.028	-5.728

Table 8. Model for the mean as function of the brood. Estimated overdispersion $\hat{\varphi}$= 2.07. Dose as a factor.

Coefficients	Value	Std. Error	*t* value
(Intercept)	1.848	0.053	34.973
Brood1	0.224	0.057	3.949
Brood2	0.118	0.029	4.047
DOSE 1	0.002	0.057	0.028
DOSE2	-0.035	0.034	-1.024
DOSE3	-0.142	0.030	-4.746
DOSE4	-0.296	0.038	-7.756

relation among offsprings will be treated as a nuisance w.r.t. the main goal of the analysis, the dose-response relationship (see Section 2.5). The choice of working correlation matrix

Table 9. Proportional odds model for ordinal response.

Coefficient	value	S.E.	Wald Z	P
$y \geq 2$	1.444	0.113	12.76	0.000
$y \geq 3$	-0.248	0.103	-2.41	0.016
x[1]	0.007	0.002	3.94	0.000
x[2]	0.000	0.001	-6.75	0.000

is based on several assumption to accommodate specific pattern of correlation amongst broods of the same subject. the different correlation structures considered are: (i) unstructured, (ii) exchangeable, (iii) Ar-1, and (iv) 1-Dependence. The detailed analysis is given in Table 10.

As before, the scale parameter shows a very high overdispersion, most likely due to the brood effect, since in the GEE setting we already adjusted for the *mother's* correlation.

All techniques considered in this section seem to agree in their conclusions and they all yield a reasonably good fitting of data. Although the conclusions of aggregated model (Bailer and Oris 1994) are in concurrence with that of disaggregated (proposed) model, yet a careful approach is needed in modeling the response variables. In aggregated model, the combined data smoothen the variability of individual broods, which as a result indicates insignificant overdispersion. However, the individual broods show very different variabilities and thus, it may be more meaningful to model them separately. Which after adjusting for overdispersion,

Table 10. GEE models for accounting for mother effect, for several structures of the working covariance matrices.

Unstructured $\hat{\phi} = 1.66$				
Coefficients	Estimate	Naive S.E.	Robust S.E.	Robust z
(Intercept)	1.904485963	0.03019642	0.03537191	53.841765
DOSE1	0.005994665	0.03666709	0.02635110	0.227492
DOSE2	-0.029931638	0.02182252	0.01306397	-2.291160
DOSE3	-0.117347458	0.01829383	0.02234431	-5.251783
DOSE4	-0.224846309	0.02038246	0.02779597	-8.089170
Brood1	0.224434659	0.06056469	0.04872245	4.606391
Brood2	0.118004693	0.02700140	0.02006532	5.881026
Exchangeable $\hat{\phi} = 1.87$				
Coefficients	Estimate	Naive S.E.	Robust S.E.	Robust z
(Intercept)	1.864278587	0.03495636	0.04224046	44.13489864
DOSE1	0.001520489	0.03883196	0.02374622	0.06403077
DOSE2	-0.033146017	0.02318780	0.01185704	-2.79547118
DOSE3	-0.136562455	0.02003663	0.02549644	-5.35613803
DOSE4	-0.275550912	0.02457207	0.03491203	-7.89272201
Brood1	0.224434659	0.06017385	0.04872245	4.60639069
Brood2	0.118004693	0.03087697	0.02006532	5.88102577
Ar-1 $\hat{\phi} = 1.99$				
Coefficients	Estimate	Naive S.E.	Robust S.E.	Robust z
(Intercept)	1.8469993943	0.04775232	0.04621282	39.967248668
Dose1	-0.0001616641	0.05152110	0.02297488	-0.007036559
DOSE2	-0.0340763106	0.03079448	0.01151979	-2.958067844
DOSE3	-0.1434067068	0.02690407	0.02709176	-5.293370066
DOSE4	-0.2961089799	0.03435477	0.03922991	-7.548041322
Brood1	0.2244346586	0.05895838	0.04872245	4.606390690
Brood2	0.1180046926	0.02875384	0.02006532	5.881025769
1-Dependence $\hat{\phi} = 1.65$				
Coefficients	Estimate	Naive S.E.	Robust S.E.	Robust z
(Intercept)	1.906927536	0.03977060	0.03873674	49.2278812
DOSE1	-0.006982931	0.04823882	0.02539908	-0.2749285
DOSE2	-0.020947724	0.02844622	0.01466632	-1.4282877
DOSE3	-0.123693886	0.02421176	0.02821871	-4.3833996
DOSE4	-0.218515028	0.02639383	0.03373749	-6.4769204
Brood1	0.224434659	0.06050157	0.04872245	4.6063907
Brood2	0.118004693	0.02112685	0.02006532	5.8810258

leads to a value of RI50, close to that of aggregated model and thus an agreement between them. In a situation of discrepancy between the two model, the proposed disaggregated model may be more reliable for it takes into account the peculiar behavior of brood one. The reason for different behavior of brood one is hard to decipher from the data. However, it could be that the dose has not taken its effect at the first reproduction stage, and therefore the brood size remains independent of the dose. Also, the small size of this brood as compared to brood two and three may be due to the lower level of maturity in mothers. Further, all the correlations structures considered in GEE analysis display similar conclusions as well. Again, the independent nature of brood one from brood two and three, makes unstructured correlation more meaningful for these data.

References

Bailer, A. J. and Oris, J. T. 1994. Assessing toxicity of pollutants in aquatic systems. In: Lange, N. Ryan, L. Billard L. Brillinger D. Conquist L., and Greenhouse, L. (eds), Case studies in Biometry. New York: Wiley and Sons. pp. 25-40.

Bishop, Y. V. V. , Fienberg S. E., and Holland, P. W. 1975. Discrete Multivariate Analysis. Cambridge, MA: MIT Press.

Carroll, R. J., and Ruppert, D. 1988. Transformations and weighting in regression. New York: Chapman and Hall.

Crawley, M. J. 1993. GLIM for Ecologists. Oxford: Blackwell Scientific Publications.

Efron, B. 1986. How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association 81(394): 461-470.

Futter, M. N. 1994. Pelagic food-web structure influences probability of mercury contamination in lake trout (*Salvelinus namaycush*). Science of Total Environment 145(1-2): 7-12.

Ganio, T. W. and Schafer, D. W. 1992. Diagnostic for Overdispersion. Journal of the American Statistical Association 87(419): 795-804.

Kaur, A., Gregori, D., Patil G. P. and Taillie, C. 1995. Ecological Applications of generalized linear models and quasi-likelihood methods: an overview. Tech. Report 95-0601. Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA, 16802, USA.

Liang, K. Y. and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.

Liang, K. Y., Zeger, S. L. and Qaqish, B. 1992. Multivariate regression analysis for categorical data. Journal of the Royal Statistical Society, Series B, 54: 3-40.

McCullagh, P. 1983. Quasi-likelihood functions. Annals of Statistics 11: 59-67.

McCullagh, P. and Nelder, J.A. 1989. Generalized Linear Models. 2 edn. London: Chapman and Hall.

Nelder, J. A. and Pregibon, D. 1987. An extended quasi-likelihood function. Biometrika 74: 221-232.

Pope, C. A., Schwartz, J. and Ransom, M. R. 1992. Daily mortality and PM10 pollution in Utah Valley. Archives of Environmental Health 47(3): 211-217.

Roy, S. N. and Kastenbaum, M. A. 1956. On the hypothesis of no interaction in a multiway contingency table. Annals of Mathematical Statistics 27: 749-757.

Schwartz, J. 1994. Air pollution and hospital admissions for the elderly in Birmingham, Alabama. American Journal of Epidemiology 139(6): 589-598.

Schwartz, J. and Levin, R. 1991. The risk of lead toxicity in homes with lead paint hazard. Environmental Research 54(1): 1-7.

Wedderburn, R. W. 1974. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika 61: 439-447.

Yule, G. U. 1912. On the methods of measuring association between two attributes (with discussion). Journal of the Royal Statistical Society 75: 579-642.