NISS

Regression Modeling of Ordinal Data with Nonzero Baselines

Minge Xie and Douglas G. Simpson

Technical Report Number 43
March, 1996

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Regression Modeling of Ordinal Data with Nonzero Baselines

Minge Xie

National Institute of Statistical Sciences

P.O.Box 14162, Research Triangle Park, North Carolina 27709, U. S. A.

Douglas G. Simpson

Department of Statistics, University of Illinois

Champaign , Illinois 61820, U. S. A.

SUMMARY

This paper develops a regression model for ordinal data with non-zero control response probabilities. The model is especially useful in dose-response studies where the spontaneous or natural response rate is nonnegligible and the dosage is logarithmic. The model generalizes Abbott's formula, which has been commonly used to model binary data with non-zero background observations. We describe a biologically plausible latent structure and develop an EM algorithm for fitting the model. The EM algorithm can be implemented using standard software for ordinal regression. Analysis of historical data on the severity of virus-induced deformities in Chicken Embryos illustrates the methodology.

Key Words: Nonzero baseline model, Ordinal data, latent structure, EM-algorithm

---

Funding for this research was provided in part by NSF-DMS 9208758 to NISS, by the US EPA Cooperative Agreement CR# 819638 with NISS and by NSF-DMS 95-05290.

1. Introduction

In their books, Finney (1972) and more recently Morgan (1992) described the analysis of binomial response data when the natural response rate is nonnegligible. In a dose-response experiment the natural response rate is often called the baseline, the background rate, or the spontaneous rate. Often death or other type of responses may occur in the control group in the absence of any stimuli as measured by the experimental variables. One of the major bioassay models for binomial observations was the binary nonzero baseline model proposed in 1925 by Abbott. Because of its simplicity and biological plausibility, Abbott's model has been commonly used to handle nonzero controls for binary data. Adapting Abbott's model to binary regression yields:

$$P(Y=1|\mathbf{x}) = \alpha + (1-\alpha)H(\mathbf{x}^T\beta), \tag{1}$$

where Y is the binary response taking possible values 0 and 1, $\mathbf{x}$ is the covariate vector, β is an unknown regression parameter vector and α, $0 \leq \alpha \leq 1$, is an unknown baseline parameter. The function H in model (1) is a general probability distribution function; its inverse function H^{-1} is called a link function. Two commonly used link functions are the logit, $H^{-1}(p) = \log\{p/(1-p)\}$, and probit, $H^{-1}(p) = \Phi^{-1}(p)$, where Φ is the standard normal distribution function.

If α is known, then Model (1) is an example of a generalized linear model as described by McCullagh and Nelder (1989). In this case the iterative weighted least squares algorithm would provides a fitting algorithm. Note the unsual feature that the response probability is bounded away from zero. Copas (1988) and Carroll and Pederson (1993) investigated closely related classes of models in which misclassifications of the response occur with specified nonzero probability.

In practice α needs to be estimated, and a more general fitting algorithm is required. Hasselblad *et al.* (1980) described an EM algorithm for the case that H is of the probit form. Barlow and Feigl (1985) suggested a variation on iterative weighted least squares for fitting a more general class of models. Preisler (1989) provided iteratively weighted least square algorithms to fit the models with probit or logistic link functions. Chapter 3 of Morgan (1992) reviewed the modeling of binary data with nonzero control responses, and suggested an EM model fitting procedure for Abbott's model. It can be shown that Morgan's EM algorithm has a smaller proportion of missing information than Hasselblad's, in situations where both apply, so Morgan's is more efficient; see Little and Rubin (1987, p.137). Here we focus on generalizability of the latent model rather than efficiency of the computing algorithm.

Our goal is to extend the Abbott-type model for analysis of ordinal data with nonzero baselines, and also to provide a convenient computational algorithm for fitting the model. We describe a Bernoulli latent structure for model (1) and generalize this latent structure to develop a non-zero baseline regression model for ordinal response data. The latent structure we describe tries to distinguish natural causes from causes due to experimental substance. In the binary case it is closely related to the latent structure implied by Morgan's (1992) analysis. Furthermore, our latent structure allows us to develop a readily implemented EM-algorithm for to obtain the parameter estimates. The E-steps calculate updated spontaneous and cumulative latent responses. The M-steps provide updated baseline parameter estimates with the updated spontaneous latent responses and get updated regression parameter estimates by maximizing standard ordinal regression models with actual responses replaced by the updated cumulative latent responses. The maximization can be handled by software which fits the ordinal regression model, and other calculations are just simple weighted proportions.

In section 2, we describe a latent structure for the Abbott type model (1), and then generalize it to obtain an ordinal non-zero baseline regression model. As in model (1), the generalized model has the flexibility to incorporate different regression schemes and techniques. The EM algorithm developed in section 3 provides a way of model fitting using standard software for multinomial data. The definition of effective dose, ED_{50}, and some technique detail of computing estimates and confidence intervals are provided in section 4. In section 5, our methodology is illustrated by an analysis of data on the severity of virus-induced deformities in Chicken Embryos. The data have been studied by McPhee *et. al.* (1984) and Morgan (1992).

2. Regression Model for Observations with Nonzero Controls

Suppose W and V are two independent random variables. They respectively have distribution,

$$W \sim \text{Bernoulli}(\alpha) \qquad V \sim \text{Bernoulli}(H) \tag{2}$$

where $H = H(\mathrm{x}^T\beta)$. Let

$$\begin{aligned} Y &= W + (1-W)V \\ &= W + V - WV, \end{aligned} \tag{3}$$

then $P(Y = 1)$ agrees with model (1). Thus any observation from model (1) can be viewed mathematically as a simple function of two latent independent Bernoulli random variables.

This latent structure has some biological explanation, for example, in risk assessment. Suppose W is the indicator of the natural (or spontaneous) mortality which is from natural causes, and V indicates the exposure (or cumulative) mortality which is caused from exposure to certain experimental substance. Formula (2) assumes that event $\{W = 1\}$ has rate α , and event $\{V = 1\}$ has rate $H(\mathbf{x}^T\beta)$, which means V itself is a binary regression. With the assumption of the independence of W and V, equation (3) leads to the statement that Y satisfies model (1). Since $\{Y = 1\}$ is equivalent to $\{W = 1\} \cup \{V = 1\}$, Y can be explained as the indicator of overall mortality, which comes from either (independent) natural causes or exposure causes.

The Bernoulli latent equation (3) can be written in other ways,

$$\mathbf{1}_{(Y=1)} = \mathbf{1}_{(W=1)} + \mathbf{1}_{(W=0)}\mathbf{1}_{(V=1)}$$

or

$$\mathbf{1}_{(Y\geq 1)} = \mathbf{1}_{(W\geq 1)} + \mathbf{1}_{(W=0)}\mathbf{1}_{(V\geq 1)}, \tag{4}$$

where $\mathbf{1}_{(C)}$ is an indicator function which equals 1 if set C is true and equals 0 if set C is false. We shall interpret equation (4) as: If the overall response Y is at least at severity level 1, then either latent spontaneous natural background response is at least at level 1, or latent spontaneous natural background response is at level 0 but latent cumulative exposure response is at least at level 1.

This argument is extended to ordinal response data as follows. Suppose that the ordered ordinal severity categories are $0, 1, \ldots, S$. If the observed overall response exceeds level s, $0 \leq s \leq S$, we believe either the latent spontaneous natural background response W exceeds level s, or W is at level l, for $0 \leq l \leq s-1$, but the latent cumulative exposure response V exceeds level $(s-l)$. If we formalize this argument, we get the extension of equation (3) or (4) to ordinal categorical cases,

$$\mathbf{1}_{(Y\geq s)} = \sum_{l=0}^{s-1} \mathbf{1}_{(W=l)}\mathbf{1}_{(V\geq s-l)} + \mathbf{1}_{(W\geq s)} \tag{5}$$

for $s = 0, 1, \ldots, S$.

To specify an ordinal model from this latent structure, we need to make assumptions on the distributions of random variables W and V. One natural choice of W is that it takes values of

$0, 1, \ldots, S$ and has distribution

$$P(W = l) = \alpha_l \quad l = 1, \ldots, S \tag{6}$$

where $\sum_{l=0}^{S} \alpha_l = 1$. We call this a multinomial trial, the polytomous analog of Bernoulli distribution. For the latent cumulative response V, parallel to binary case, we often assume it is an regression on an explanatory variable $\mathbf{x}$. That is, we assume V takes values of $0, 1, \ldots, S$ and has a distribution with the form of an ordinal regression model,

$$\begin{aligned} P(V \geq s|\mathbf{x}) &= \mathcal{H}(s, \mathbf{x}), \quad \text{if } s = 1, \ldots, S; \\ &= 1 \quad \text{if } s = 0, \end{aligned} \tag{7}$$

where $\mathbf{x}$ is the regression covariates and $\mathcal{H}(s, \mathbf{x})$ is a general function with $0 \leq \mathcal{H}(s, \mathbf{x}) \leq 1$ and $\mathcal{H}(s, \mathbf{x})$ decreases in s. In general, $\mathcal{H}$ could be any type of regression model, linear, non-linear or even non-parametric. As long as there is a procedure to fit the regression model, the choice of $\mathcal{H}$ would not affect the EM fitting procedure which will be presented in the next section. In this paper, we only focus our attention on V in the class of generalized linear models.

One special generalized linear regression model of (7), which is a popular choice among ordinal regression models, can be written as,

$$\begin{aligned} P(V \geq s|\mathbf{x}) &= H(\gamma_s + \mathbf{x}^T \beta) \quad \text{if } s = 1, \ldots, S; \\ &= 1 \quad \text{if } s = 0, \end{aligned} \tag{8}$$

where $(\gamma_1, \ldots, \gamma_S, \beta)$ are regression parameters and H^{-1} is a link function. If H^{-1} is a logit link, model (8) is referred as the proportional odds model (McCullagh, 1980).

Assuming W and V are independent, based on equations (5), (6) and (7), we propose a nonzero baseline ordinal regression model,

$$P(Y \geq s|\mathbf{x}) = \sum_{l=0}^{s-1} \alpha_l \mathcal{H}(s - l, \mathbf{x}) + \sum_{l=s}^{S} \alpha_l, \tag{9}$$

for $s = 0, 1, \ldots, S$.

If $S = 2$, this model reduces to model (1). If $s = 0$, the first summation vanishes and the probability equals to 1, the sum of all α's. To accommodate some special situations, we sometimes allow some of the baseline rates α to be set equal to zero. For instance, if the maximum background severity category level is K with $K < S$, we shall let $\alpha_{K+1} = \alpha_{K+2} = \ldots = \alpha_S = 0$ and drop them out from the model. In model (9), α_l is explained as the spontaneous natural rate of response at severity level $l = 0, 1, \ldots, S$ and the regression parameters only appear in the generalized linear regression part $\mathcal{H}$.

As model (1), model (9) has some advantage to deal with non-zero control responses particular when some explanatory regression variable is on a logarithmic scale. For example, in dose-response models, with the dose variable entering the model on log-scale, the cumulative regression parts disappears for control responses. It implies that control responses only come from spontaneous background causes.

3. EM Algorithm

Model (9) provides formulas of response rates for ordinal observations from multinomial trials. In practice, we are more likely to have observations $y_i = (y_{i0}, y_{i1}, \ldots, y_{iS})$, $i = 1, \ldots, k$, that are independent samples from multinomial distribution, Multinomial$(n_i; p_{i0}, p_{i1}, \ldots, p_{iS})$, where rates $\{p_{is}, s = 0, 1, \ldots, S\}$ satisfy the condition that $\sum_{l=s}^{S} p_{il}$ follows the form of (9). In a dose-response model, for instance, the y_{is} is often the number of responses occurred at severity level s, $s = 0, \ldots, S$, when n_i individuals are studied at dosage d_i, $i = 1, \ldots, k$. If we denote $\mathcal{U}_{i,s} = \mathcal{H}(s, \mathbf{x}_i) - \mathcal{H}(s+1, \mathbf{x}_i)$, and let θ represent all parameters appeared in the model, then the log-likelihood function of observations $\{y_i, i = 1, \ldots, k\}$ will be

$$l(\theta|y) = \sum_{i=1}^{n} \sum_{s=0}^{S} y_{is} \log p_{is},$$

where $p_{is} = \sum_{l=0}^{s} \alpha_l \mathcal{U}_{i,(s-l)}$ for $s = 0, 1, \ldots, S-1$ and $p_{iS} = \sum_{l=0}^{S} \alpha_l \mathcal{H}(S-l, \mathbf{x})$. Because direct maximization of $l(\theta|y)$ is tedious in general, special effort is required to obtain the parameter estimates. Based on the latent structure discussed in the last section, we introduce a natural EM approach to fit nonzero baseline model (9) for multinomial data.

Suppose $\{Y_{ij}^*|i = 1, \ldots, k, j = 1, \ldots, n_i\}$ is a latent realization of $Y_i = (Y_{i0}, Y_{i1}, \ldots, Y_{iS})$, that is, for each i, $i = 1, \ldots, k$, Y_{ij}^*, $j = 1, \ldots, n_i$, are independent multinomial trials from Multinomial$(1; p_{i0}, p_{i1}, \ldots, p_{iS})$, and $Y_{is} = \sum_{j=1}^{n_i} \mathbf{1}_{(Y_{ij}^*=s)}$, where $\mathbf{1}_{(C)}$ is an index function. For $j = 1, \ldots, n_i$, $i = 1, \ldots, k$, we choose a pair of latent variables (W_{ij}^*, V_{ij}^*) as multinomial trials such that their rates satisfy the form of (7) and (8) respectively, and together with Y_{ij}^*, they satisfy equation (5). Write $W_{is} = \sum_{j=1}^{n_i} \mathbf{1}_{(W_{ij}^*=s)}$, $V_{is} = \sum_{j=1}^{n_i} \mathbf{1}_{(V_{ij}^*=s)}$ and denote $W_i = (W_{i0}, \dots, W_{iS})$, $V_i = (V_{i0}, \dots, V_{iS})$, then (W_i, V_i) can be viewed as a pair of latent random variables for Y_i. To derive an EM algorithm, we view $\{(w_i, v_i); i = 1, \dots, k\}$ as the full data. The joint log-likelihood function of $\{(w_i, v_i), i = 1, \dots, n\}$ is,

$$l(\theta|(w_i, v_i), i = 1, \dots, n) = \sum_{i=1}^{n}\sum_{l=0}^{S} w_{il} \log \alpha_l + \sum_{i=1}^{n}\sum_{s=0}^{S} v_{is} \log \mathcal{U}_{i,s}.$$

After some standard mathematical calculations (see appendix for some detail), we find that the EM algorithm iterates through the following two steps:

E-step:

$$\begin{aligned} c_{i,l}^{(m)} &= E(W_{il}|Y_{ir} = y_{ir}, \alpha^{(m)}, \mathcal{U}_{i,r}^{(m)}, r = 0, \dots, S) \\ &= \sum_{t=l}^{S-1} y_{it} \frac{\alpha_l^{(m)} \mathcal{U}_{i,(t-l)}^{(m)}}{\sum_{l'=0}^{t} \alpha_{l'}^{(m)} \mathcal{U}_{i,(t-l')}^{(m)}} + y_{iS} \frac{\alpha_l^{(m)} \mathcal{H}_{i,(S-l)}^{(m)}}{\sum_{l'=0}^{S} \alpha_{l'}^{(m)} \mathcal{H}_{i,(S-l')}^{(m)}} \end{aligned}$$

$$\begin{aligned} d_{i,s}^{(m)} &= E(V_{is}|Y_{ir} = y_{ir}, \alpha^{(m)}, \mathcal{U}_{i,r}^{(m)}, r = 0, \dots, S) \\ &= \sum_{t=s}^{S-1} y_{it} \frac{\alpha_{(t-s)}^{(m)} \mathcal{U}_{i,s}^{(m)}}{\sum_{s'=0}^{t} \alpha_{(t-s')}^{(m)} \mathcal{U}_{i,s'}^{(m)}} + y_{iS} \frac{\alpha_{(S-s)}^{(m)} \mathcal{H}_{i,s}^{(m)}}{\sum_{s'=0}^{S} \alpha_{(S-s')}^{(m)} \mathcal{H}_{i,s'}^{(m)}} \end{aligned}$$

M-step:

$$\alpha_l^{(m+1)} = \frac{1}{N} \sum_{i=1}^{n} c_{i,l}^{(m)}$$

$$\beta^{(m+1)} = \text{argmax}_\beta \, \{\sum_{i=1}^{n}\sum_{s=0}^{S} d_{i,s}^{(m)} \log \mathcal{U}_{i,s}\}$$

where $N = \sum_{i=1}^{k} n_i$. The last maximization is exactly the maximization of loglikelihood function of usual ordinal regression model (7) with true observations replaced by $d_{i,s}^{(m)}$. Any standard software which fits the ordinal regression models can be used to handle this step. For instance, we can use the SAS (SAS Institute, Inc.) PROC LOGISTIC to fit ordinal regression model for several choices of $\mathcal{H}$.

In the binary model (1), $\alpha = \alpha_1 = 1 - \alpha_0$, and $\mathcal{H}(s, \mathbf{x})$ is reduced simply to $H(\mathbf{x}^T\beta)$. The EM algorithm becomes

E-step:

$$c_{i,1}^{(m)} = E(W_{i1}|Y_i = (y_{i0}, y_{i1}), \alpha^{(m)}, H^{(m)}) = y_{i1} \frac{\alpha^{(m)}}{(1 - \alpha^{(m)})H_i^{(m)} + \alpha^{(m)}}$$

$$d_{i,1}^{(m)} = E(V_i|Y_i = (y_{i0}, y_{i1}), \alpha^{(m)}, H^{(m)}) = y_{i1}\frac{H_i^{(m)}}{\alpha^{(m)}(1 - H_i^{(m)}) + H_i^{(m)}}.$$

M-step:

$$\alpha^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} c_{i,1}^{(m)},$$

and $\beta^{(m+1)}$ is the solution of equation:

$$\sum_{i=1}^{k} n_i(\frac{d_{i,1}^{(m)}}{n_i} - H(\mathbf{x}_i^T\beta))\mathbf{x}_i = 0.$$

This equation can be solved via standard statistical software for logistic regression and its variants.

Apparently the E-steps calculate the expectations of updated spontaneous and cumulative latent variables, the M-steps provide the updated baseline paramenter estimates by taking average of updated spontaneous latent observations and obtain the updated regression parameter estimates by solving a maximization problem which only involves updated cumulative latent observations. Except for the maximization, there are only simple calculations. These EM-algorithms could easily be implemented in a standard software which provides a way to solve this type of maximization problems.

In models (1) and (9), the corresponding likelihood functions could possibly have multiple modes. The EM algorithms are not guaranteed to converge to a global maximum; see, Dempster *et al.*, 1977 and Wu, 1983. However, the projected log-likelihood $Q(\theta^*|\theta) := E\{l(\theta^*|(w_i, v_i), i = 1, \ldots, n)|\theta, y_i, i = 1, \ldots, n\}$ is continuous in both θ^* and θ. By theorem 2 of Wu (1983), EM algorithms presented above are guaranteed to converge to a stationary point. If one would like to verify this stationary point is a local maximum or not, numerical calculation of the second derivative matrix are needed. This numerical calculation could be obtained, for example, by techniques presented in Meng and Rubin (1991).

To find appropriate estimates, we need reasonable starting values for the EM algorithms. One natural choice is to use the observed proportions of each severity categories of control responses as starting values for the spontaneous baseline rates. To get starting values for the regression parameters, we fit the set of non-control observations or a set of simply adjusted exposure-only observations to a standard ordinal regression model. The method of "adjusted" means to calculate the suitable proportions of non-control observations through model (9) by treating the baseline parameters as they were known (set equal to the proportion of control responses). In binary case, this method of adjusting is discussed in Barlow and Feigl (1985), Hoekstra (1987) and Morgan (1992).

As discussed in section 2, some of the baseline parameters in model (9) may have constraint of being zero and could be dropped out from the original full model. In this situation, the EM algorithms obtained above still work. By setting corresponding baseline start point to be zero, we can ensure the constrained baseline parameters to be zero through out the whole EM procedure. In fact, if each group size n_i, $i = 1, \ldots, k$, is large enough, a stronger statement can be obtained from large sample theory: the estimates of the zero constrained parameters obtained from our EM algorithm are always zero, independent of the choice of the starting value.

4. Effective Concentration Parameters

The *effective dose*, ED_{50}, at severity level s, $s = 1, \ldots, S$, is defined as the concentration that produces a parallel degree of 50% responses that exceed severity level s (e.g., death plus deformities or just death). In model (9), if the independent regression variables $\mathbf{x}$ include $\log_{10}$(dosage) and some other covariates $\tilde{\mathbf{x}}$, then the ED_{50} at severity level s solves the equation,

$$Q_s\{[\tilde{\mathbf{x}}, \log_{10}(\mathrm{ED}_{50})]; \theta\} = 0.50,$$

where $Q_s(\mathbf{x}; \theta)$ is the right hand side of model (9). As ED_{50} is a function of θ, its estimate can be obtained by substitution of $\hat{\theta}$ for θ. Except for certain special cases, one needs an iterative algorithm to solve for ED_{50} values. For instance, we can use Newton-Raphson algorithm, since the evaluation of $Q_s(\mathbf{x}; \theta)$ and its first and second derivative is often fairly easy.

There are several ways to obtain confidence intervals for ED_{50}. Three classical approaches, which are asymptotically equivalent, are Wald type approach, likelihood ratio based approach (Cox and Hinkley, 1974) and the so-called fiducial limits of Fieller (1954). In their books, Finney (1978) and Morgan (1992) carefully accounted the Fieller type confidence interval. Alho and Valtonen (1995) provided a Newton algorithm to calculate the endpoints of the likelihood ratio based confidence intervals and provided simulation comparison between likelihood ratio based confidence intervals and Fieller type fiducial limits. The Wald type confidence intervals are often provide wider confidence interval than the other two, and because of their lack of invariance, the performance may be enhanced by a judicious transformation of the parameter estimates. In our analysis of the Chicken Embryo data in the next section we adopt a modified Wald approach in which we compute the

Wald confidence interval for the logarithmic ED_{50}, and then exponentiate the endpoints. These Wald type confidence intervals appear adequate to address our point.

Supposing temporarily that the asymptotic variance-covariance matrix of $\hat{\theta}$, say $\mathbf{V}$, is known, we can obtain the Wald type confidence interval for ED_{50} by the delta method as follows. Denote $\log_{10}(\mathrm{ED}_{50})$, the $\log_{10}$ 50% effective dose at severity level s, by $G_{0.5,s}(\theta)$, and write

$$\mathbf{a} = \{\frac{\partial}{\partial\theta}G_{0.5,s}(\theta)\}|_{\theta=\hat{\theta}} = [\{\frac{\partial}{\partial x}Q_s(x;\theta)\}|_{\theta=\hat{\theta}}]^{-1}\{\frac{\partial}{\partial\theta}Q_s(x;\theta)\}|_{\theta=\hat{\theta}}$$

By the delta method, the variance estimate of $\log_{10}(\mathrm{ED}_{50})$ is $\mathbf{a}^T\mathbf{V}\mathbf{a}$. From this we obtain the Wald type normal confidence interval for $\log_{10}(\mathrm{ED}_{50})$, and exponentiate its endpoints to obtain the confidence interval for ED_{50}. Recall that the Wald type confidence interval is transformation dependent. The use of back-transformed technique here to obtain confidence interval for ED_{50} is based on our experience that the distribution of the estimate of log-scaled ED_{50} is more close to normal than the original unscaled one. In addition, this back-transformed method ensures that the confidence interval of ED_{50} satisfies the positivity condition.

As we know, ordinary EM-algorithm usually will not produce covariance matrix estimate for parameters. To obtain the asymptotic covariance matrix $\mathbf{V}$, some special treatment is needed. Approaches to obtain covariance matrix $\mathbf{V}$ through EM-algorithm have been suggested, for example, by Louis (1982), Meilijson (1989) and Meng and Rubin (1991). These approaches are mostly useful under the situation where the direct computation of the log-likelihood for observed data and its up to second order derivatives is difficulty. But they are often slow and usually needs extra program effort. In some cases, for example in binary model (1) and some special case of model (9), log-likelihood function and its derivatives can be calculated directly. In these cases, we suggest to evaluate covariance matrix through observed information matrix or obtain variance estimate through likelihood ratio test statistics.

5. Example: The Analysis of Chicken Embryo Data

5.1 The Data

The data of Table 1 form a subset of data from Jarrett *et al.* (1981), which were quoted in Table 1.7 of Morgan (1992, page 10). The objective was to investigate the effects of arboviruses injected into chicken embryos and to qualify the potency of arboviruses, with a view ultimately to assess how these might affect lamb fetuses (see, Morgan, Example 1.5, Page 9). In the experiments, eggs were inoculated with a range of virus and several inoculum levels and candled daily for 14 days to check viability. The surviving embryos were then examed for gross abnormalities and their results were reported 4 days later; see Jarrett *et al.* (1981) and McPhee *et al.* (1984) for more detail. The resulting data of a control group and two arboviruses, Facey's Paddock virus and Tinaroo virus, are given in Table 1. There are three levels of possible responses – death, alive but deformed, and alive but not deformed. The need to examine the dependence of their responses on the amount of injected viruses leads to a typical statistical ordinal regression analysis.

Table 1 (Morgan, 1992) *The effect of two arboviruses on chicken embryo*

	Inoculum titre	number of	number of	number of	total
Virus	(PFU/egg)	Death	deformed	not deformed	number of eggs
Facey's Paddock	3	3	1	13	17
	18	4	1	14	19
	30	8	2	9	19
	90	17	1	2	2
Tinaroo	3	1	0	18	19
	20	2	0	17	19
	2,400	4	9	2	15
	88,000	9	10	0	19
Control		1	0	17	18

Jarrett *et al.* (1981), McPhee *et al.* (1984) and Morgan (1992) studied the arbovirus data two of which are presented in Table 1. Because the control rate of death is "small" (at level of 5.88% in Table 1 and 7.7% for the control group in McPhee *et al.*, 1984), and the control rate of deformed is zero, they ignored the control group observations and used standard proportional odds model to analyze the data. The primary covariate variable they used was log-dosage. Morgan (1992) and Jarrett *et al.* (1981) indicated the failure of model fitting for Tinaroo virus data. In McPhee *et al.* (1984), they provided detail experiment and model fitting information, and listed estimates of ED_{50}'s and their confidence intervals for different viruses.

5.2 The Model

In model (9), the choice of linear regression model for the analysis of the dependence of the latent cumulative exposure effect on explanatory variables depends on the data and the purpose of study. For the analysis of chicken embryo data, we select the proportional odds model, i.e., regression model (8) with logistic link. One of the reasons is that Morgan (1992), Jarrett *et al.* (1981) and McPhee *et al.* (1984) all used this model. The difference is our model also includes baseline analysis, and allows us to use the control information.

Since the observed number of deformed embryos in control group is zero, the fitting of model (9) to the chicken embryo data provide us zero as the estimate of baseline parameter for deformed category. Notice zero is at the boundary of a baseline parameter space. The first derivative of the log- likelihood function with respect to the baseline parameter for deformed category is not zero at the estimated point, and the standard large sample theories is no longer valid. Some results involving complicated boundary asymptotics, for example, discussed in Chernoff (1951) and Self and Liang (1987) may be required. Instead, to avoid complications, we assume the spontaneous occurence of nonlethal deformities is zero. One can argue that the true background deformed rate is zero or close to zero and it is ignorable. In the goodness-of-fit test the null hypothesis is Model (9) with a constraint which restricts the baseline parameter for deformed category to be zero. Under this assumption regular χ^2 theory applies, because all parameter estimates are in the interior of the parameter space.

With this simplification, we denote the baseline parameter of the lethal category as α, and the baseline parameter for the live/non-deformed category is $1 - \alpha$. The incidence rates associated with severity level not deformed, deformed and lethal for non-control data will be p_{i1}, $1 - p_{i1}$ and $p_{i1} - p_{i2}$ respectively, where

$$p_{i1} = (1-\alpha)H(\gamma_1 + \mathbf{x}_i^T \beta) + \alpha, \tag{10}$$

$$p_{i2} = (1-\alpha)H(\gamma_2 + \mathbf{x}_i^T \beta) + \alpha. \tag{11}$$

In (11) and (12), the i represents all dosage groups except the control group, H is the logistic link function, and γ_1 and γ_2 are intercept parameters which are corresponding to severity categories deformed and death respectively (their zero reference is the intercept corresponding to severity level not deformed). We use $y_{cj}, j = 0, 1, 2$ to represent the observations of the control group associated with severity response not deformed, deformed and death, and write $\theta = (\alpha, \gamma_1, \gamma_2, \beta)^{\mathrm{T}}$. The likelihood function of the observed data is,

$$\mathcal{L}(\theta) = (1-\alpha)^{y_{c0}} \alpha^{y_{c2}} \prod_{i=1}^{k} (1 - p_{i1})^{y_{i0}} (p_{i1} - p_{i2})^{y_{i1}} p_{i2}^{y_{i2}} \mathbf{1}_{(y_{c1}=0)}.$$

Because observed $y_{c1} = 0$, the loglikelihood function exists,

$$l(\theta|y) = y_{c0}\log(1-\alpha) + y_{c2}\log(\alpha) + \sum_{i=1}^{k}\{y_{i0}\log(1-p_{i1}) + y_{i1}\log(p_{i1}-p_{i2}) + y_{i2}\log(p_{i2})\}. \quad (12)$$

The EM algorithm presented in Section 3 is used to find estimates from this log-likelihood function. To implement constraint $\alpha_1 = 0$, we set $\alpha_1^{(0)}$ equal to zero, which ensured that $\alpha_1^{(1)}$, $\alpha_1^{(2)}$, ...) were are all zero.

It is worth mentioning that, in this example, models (10) and (11) are both nonzero baseline binary models. This is not true if the baseline parameter for deformed category is non-zero. Furthermore, even if the incidence rates in models (10) and (11) have the formula of two binary models, the model differs from an approach which directly combines the results of fitting two separate binary models. Our likelihood approach analyzes (10) and (11) simultaneously.

5.3 Model Fitting Results and Conclusions

From Table 1, we form three data sets: (i) observations from experiment using Tinaroo virus plus control data; (ii) observations from experiment using Facey's Paddock virus plus control data; and (iii) the whole data (Tinaroo observations plus Facey's Paddock observations plus control observations). Since the experiments of both viruses share a same control experiment, analyzing data set (iii), i.e., total observations altogether is more appropriate than analyzing data set (i) or (ii) individually. Because all previous work by Jarrett *et al.* (1981), McPhee *et al.* (1984) and Morgan (1992) treated those two virus data separately, we will also analysis data sets (i) and (ii) for the purpose of strict comparison.

The calculations were carried out in the S-plus (MathSoft, Inc.) environment. In S-plus, there are no standard functions to perform the model fitting task for proportional odds models. But we have written S-plus based software to fit standard ordinal regression models (see, Simpson *et al.* 1995). We use this customized program to perform the maximization and coded our EM algorithm in S-plus. The computer programs are available upon request.

Table 2 presents the model fitting results from our EM code. We calculate the parameter estimates and deviances while we fit our model to the three data sets (i), (ii) and (iii) formed from Table 1. The deviances are 4.90, 4.29 and 8.59 respectively. Under the models with zero constraint on deformed category baseline, the degree of freedom for the goodness of fit test of these three deviances are 5, 5 and 10. All three deviances are smaller than their corresponding χ^2 95% critical values. These indicate that all three data sets fit associated constrained models well. We also repeated model fitting done by Jarrett *et al.* (1981), McPhee *et al.* (1984) and Morgan (1992) and tried standard proportional odds model both to data (ii) and to data from experiment of injecting Tinaroo virus (data (ii) without control) with dose entering model at original scale. All fittings provide large deviances and prove to be lack of fit. There is a trend of natural causes in experiment data of using Tinaroo virus, and only nonzero baseline model have picked it up. Nonzero baseline model solves the poor model fitting problem of Tinaroo data reported in Morgan (1992) and Jarrett *et al.* (1981).

Table 2 *Parameter Estimates for Jarrett's Chicken Embryo Data:*

	Baseline		Regress					
Virus	Not- deform	Death	Intrcpt$_T$ Deform	Intcpt$_T$ Death	Slope$_T$ Dose	Intrcpt$_T$ Deform	Intcpt$_T$ Death	Slope$_T$ Dose
Cntrl& Tinaroo	0.89	0.11	-9.97	-17.09	3.39			
	(0.74, 0.96)	(0.04, 0.26)	(-12.24, -7.70)	(-19.84, -14.35)	(2.86,3.92)			
Cntrl& Paddock	0.93	0.07				-4.41	-4.84	3.05
	(0.75,0.98)	(0.02,0.25)				(-5.06,-3.76)	(-5.51,-4.18)	(2.48,3.62)
Cntrl& Both	0.89	0.11	-9.70	-16.66	3.30	-5.23	-5.72	3.52
	(0.76,0.95)	(0.05,0.24)	(-11.89, -7.52)	(-19.34, -13.98)	(2.80, 3.81)	(-5.92, -4.55)	(-6.42, -5.02)	(2.92,4.12)

We computed the covariance matrix **V** by calculate the second derivative of the loglikelihood function (12). The Wald type 95% confidence interval for model parameters are reported in the parenthesis of Table 1. Due to the same reason we described for confidence interval calculation of ED$_{50}$, the confidence intervals of baseline parameters are obtained by a back-transformed technique as well. The parameter under consideration here is $\log(\alpha/(1-\alpha))$.

According to the definition of ED$_{50}$ and the discussion in section 4, we calculated estimates of ED$_{50}$'s for death and for death plus deformed and their Wald type 95% confidence intervals. The numerical results are reported in Table 3. Also in Table 3, we cited the results of these two viruses posted in McPhee *et al.* (1984). The confidence interval reported in McPhee *et al.* (1984) are Fieller type fiducial limits.

Table 3 ED_{50} *Comparison for Jarrett's Chicken Embryo Data*:

		Death			Death&Deformed		
Virus	Model	ED_{50}	95%-Low	95%-Up	ED_{50}	95%-Low	95%-Up
Fancey's Paddock	Baseline (All Data)	39.1	18.2	84.2	28.4	15.8	51.2
	Baseline (Cntrl&Paddock)	36.7	16.1	83.5	26.5	13.7	51.5
	McPhee's (Paddock only)	32.7	18.7*	65.5*	22.8	11.9*	40.8*
Tinaroo	Baseline (All Data)	102,149.6	48,688.7	214,311.0	800.3	284.6	2249.9
	Baseline (Cntrl&Tinaroo)	102,615.8	49,887.0	211,077.1	813.9	293.4	2257.9
	McPhee's (Tinaroo only)	41,000	10,200*	244,000*	465	116*	1918*

Remark: In McPhee et al. (1984), they separately fitted Paddock and Tinaroo (no control) data to proportional odds model. The confidence intervals reported in their method were otained by using Fieller's theorem (see, Finney, 1971)

Although we do not have formal theories available to compare the model fitting results of data sets (i) or (ii) with those of data set (iii), numerically they are close. The confidence intervals obtained from data (iii) are slightly tighter, since more information (observations) are used altogether to estimate these limits. We find that both ED_{50}'s of injecting Facey's Paddock virus reported in McPhee *et al.* (1984) are slightly lower than what we obtained, but well within the confidence intervals. This could be because the nonzero baseline rates is small (7% in our model for Paddock data plus control, empirically proportion 5.88% from Table 1) and plus single Facey's Paddock virus data fits standard proportional odds model well (Jarrett *et al.*, 1981 and Morgan, 1992). However, ED_{50} of death level for Tinaroo virus are fallen out of our corresponding confidence interval from non-zero baseline models. We tend to believe this is caused by the lack-of-fit of the single Tinaroo data to the simple proportional regression model.

Correction by Abbott type model (1) for binomial data is sometimes not considered necessary if control mortality is slight, say less than 5 or 10%; see Morgan, 1992, page 95, American Public Health Association *et al.*, 1981. One could have similar claim for ordinal observations if standard ordinal (zero baseline) regression fits the data well. The model fitting results for Facey's Paddock data supported this claim. However, the analysis of Tinaroo data provide evidence that even the "small" baseline rate can effect substantially the model fitting results. Simply ignoring control responses could lead to inefficiency, and it makes goodness of fit testing problematic.

6. Discussion

In this paper, we describe a biologically plausible latent structure for data with natural cause responses. We build an ordinal non-zero baseline regression model and develop associated EM algorithms. The latent structure plays an important roll. The strength of the latent structure is that it separates cumulative exposure causes from spontaneous natural causes. This separation not only helps the development of the ordinal nonzero baseline model, it also makes incorporating standard ordinal model fitting softwares into our EM schemes possible, and therefore makes the implementation easy.

One further generalization of model (1) and model (9) is to multiple ordinal responses. Based on the same latent idea, we can obtain a reasonable multivariate nonzero baseline model for binary and ordinal observations, and we can derive similar EM-algorithms. Also, we can incorporate marginal and generalized likelihood analysis into these nonzero baseline models. Some discussions of multivariate nonzero baseline binary model and marginal analysis can be found in Xie (1996, section 3, Chapter 5, Ph.D. dissertation, University of Illinois at Urbana-Champaign).

Our model fitting of the chicken embryo data of Table 1 indicates the practical use of our nonzero baseline model. It is especially useful when the natural background rate is not ignorable. In this paper we emphasized regression modeling. The latent idea and nonzero baseline models can also apply to contingency table where the response levels are ordered.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Gary Foureman of U.S. EPA for his helpful discussion and support of this research.

Appendix

Calculation of Conditional Expectation in Section 4

The difficult part of the derivation of EM algorithm is to obtain the conditional expectation formula for the latent variable W's and V's conditional on the observed Y. The following is some high light of the calculation.

$$\begin{aligned}
E(W_{il}|Y_{i1}^* = y_{i1}^*, \ldots, Y_{in_i}^* = y_{in_i}^*) &= E(\sum_{j=1}^{n_i} \mathbf{1}_{(W_{ij}^*=l)}|Y_{i1}^* = y_{i1}^*, \ldots, Y_{in_i}^* = y_{in_i}^*) \\
&= \sum_{j=1}^{n_i} P(W_{ij}^* = l|Y_{i1}^*, \ldots, Y_{in_i}^*) \\
&= \sum_{j=1}^{n_i} \frac{P(Y_{ij}^* = y_{ij}^*, W_{ij}^* = l)}{P(Y_{ij}^* = y_{ij}^*)} \\
&= \sum_{j=1}^{n_i} \{ \sum_{m=l}^{S-1} \frac{\alpha_l \mathcal{U}_{i,m-l}}{\sum_{l'=0}^{m} \alpha_{l'} \mathcal{U}_{i,m-l'}} \mathbf{1}_{(y_{ij}^*=m)} + \frac{\alpha_l \mathcal{H}_{i,S-l}}{\sum_{l'=0}^{S} \alpha_{l'} \mathcal{H}_{i,S-l'}} \mathbf{1}_{(y_{ij}^*=S)} \} \\
&= \sum_{m=l}^{S-1} \frac{\alpha_l \mathcal{U}_{i,m-l}}{\sum_{l'=0}^{m} \alpha_{l'} \mathcal{U}_{i,m-l'}} y_{il} + \frac{\alpha_l \mathcal{H}_{i,S-l}}{\sum_{l'=0}^{S} \alpha_{l'} \mathcal{H}_{i,S-l'}} y_{iS}
\end{aligned}$$

So,

$$\begin{aligned}
E(W_{il}|Y_{i0} = y_{i0}, Y_{i1} = y_{i1}, \ldots, Y_{iS} = y_{is}) &= E(E(W_{il}|Y_{i1}^*, \ldots, Y_{in_i}^*)|Y_{i0} = y_{i0}, Y_{i1} = y_{i1}, \ldots, Y_{iS} = y_{is}) \\
&= \sum_{m=l}^{S-1} \frac{\alpha_l \mathcal{U}_{i,m-l}}{\sum_{l'=0}^{m} \alpha_{l'} \mathcal{U}_{i,m-l'}} y_{il} + \frac{\alpha_l \mathcal{H}_{i,S-l}}{\sum_{l'=0}^{S} \alpha_{l'} \mathcal{H}_{i,S-l'}} y_{iS}
\end{aligned}$$

The calculation of $E(V_{il}|Y_{i0} = y_{i0}, Y_{i1} = y_{i1}, \ldots, Y_{iS} = y_{is})$ is similar.

References

Abbott, W.S. (1925). A method of computing the effectiveness of an insecticide. *J. Econ. Entomol.*, 18, 265-267.

Alho, J.M. and Valtonen, E. (1995). Interval Estimation of Inverse Dose-Response. *Biometrics*, 51, 491-501.

Barlow, W.E.M and Feigl, P.(1985). Analyzing binomial data with a nonzero baseline using GLIM. *Computational Statistics and Data Analysis*, 3, 201-204.

Carroll, R.J. and Pederson, S. (1993). On robustness in the logistic regression model. *J.R. Statist. Soc.* B, 55, 693-706.

Copas, J.B. (1988). Binary regression models for contaminated data (with discussion). *J.R. Statist. Soc.* B, 50, 225-265.

Cox, C. (1984). Generalized linear models - The missing link. *Appl. Statist.* 33, 18-24.

Dempster, A.P., Laird, N.M. and Rubin, D.B.(1977). Maximum likelihood from incomplete observations. *J. Roy. Statist. Soc. B*, 39, 1-38.

Fieller, E.C. (1954). Some problems in interval estimation. *Journal of Royal Statistical Society, Series B*, 16, 175-185.

Finney, D.J. (1971). *Probit analysis*, 3rd ed., Cambridge University Press, Cambridge.

Hasselblad, V., Stead A.G. and Creason, J.P. (1980). Multiple probit analysis with a nonzero background. *Biometrics*, 36, 659-663.

Hoekstra, J.A. (1987). Acute bioassays with control mortality. *Water, Air and Soil Pollution*, 35, 311-317.

Jarrett, R.G., Morgan, B.J.T. and Liow, S. (1981). The effects of viruses on death and deformity rates in chicken eggs. *consulting report VT 81/37.* CSIRO Division of Mathematics and statistics, Melbourne.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data.* New York: Wiley.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, B*, 44, 226 - 233.

McPhee, D.A., Parsonson, I.M., Della-Porta, A.J. and Jarrett, R.G. (1984). Teratogenicity of Australian simbu serogroup and some other bunyaviridae viruses: the embryonated chicken egg as a model. *Infection and Immunity*, 43, 413-420.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models* (Second edition). London: Chapman and Hall.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, Ser. B*, 51, 127 - 138.

Meng, X.-L. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86, 899-909.

Morgan, B.J.T. (1992). *Analysis of quantal response data.* London: Chapman and Hall.

Preisler, H.K., (1989). Fitting dose-response data with non-zero background within generalized linear and generalized additive models. *Comp. Stat. & Data Anal.*, 7, 279-290.

Simpson. D.G, Carroll, R.J., Schmedieche, H. and Xie, M. (1995). Documentation for categorical regression risk assessment (CatReg). Report to U.S.EPA, National Center for Environmental Assessment, Research Triangle Park, North Carolina.

Wu, C.-F. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11, 95-103.