Mixed Logistic Regression Models

Peiming WANG and Martin L. PUTERMAN

This article studies binomial mixture models that include covariates in binomial parameters and mixing probabilities. This model contains logistic regression, nonparametric mixed logistic regression (Follmann and Lambert 1989) and independent binomial mixture models as special cases, and provides an alternative to quasi-likelihood and beta-binomial regression for modeling extra-binomial variation. Estimation methods based on the EM and quasi-Newton algorithms, properties of these estimates, a model selection procedure, residual analysis, and goodness of fit are discussed. This methodology is motivated and illustrated with an example. A Monte Carlo study investigates behavior of the estimates and model selection criteria.

Key Words: Binomial data; EM algorithm; Extra-binomial variation; Mixture models; Residual analysis.

1. Introduction

Logistic regression has been widely used for analyzing binomial data in which each observation consists of a finite valued response variable and a vector of covariates or predictors. Areas of applications include epidemiology, quantal bioassay, and the social sciences. Sometimes there may be more variation than can be attributable to the usual binomial mean-variance relationship, suggesting the need for alternative models. In this article, we propose a class of finite mixture models that model this extra-binomial variation by including covariates in the binomial parameters, the mixing probabilities, or both. In some applications, the proposed model may provide an interpretable alternative to other approaches for dealing with extra-binomial variation.

Several alternatives have been proposed to the binomial distribution to allow for extra-binomial variation. Beta-binomial models (e.g., Crowder 1978; Otake and Prentice 1984; Kahn and Raftery 1996) assume that the binomial parameter follows a beta distribution. The double-binomial distribution (Efron 1986) modifies the binomial distribution through rescaling its binomial denominator and allows covariates in the mean and variance. Quasi-likelihood models (McCullagh and Nelder 1989 and Williams, 1982) modify the binomial distribution by specifying only the first two moment functions rather than the complete distribution. Random effect models include an additive random effect in

the regression component: specifying only the first two moments of the random variable yields a quasi-likelihood model (e.g., Williams 1982); specifying a distribution for the random variable, for example, normal, yields a compound binomial model (Pierce and Sands 1975; Anderson and Aitkin 1985; Breslow and Clayton 1993); and allowing an arbitrary distribution corresponds to the nonparametric mixed binomial model of Follmann and Lambert (1989). The approach taken in this article is closest in spirit to that of Follmann and Lambert but extends it in several ways. A similar model for extra-Poisson variation has been studied extensively in Wang (1994), Wang, Puterman, Le, and Cockburn (1996) and Wang, Cockburn, and Puterman (1998).

Cox (1983) discussed the effects of overdispersion (extra-binomial variation) on the consistency and efficiency of maximum likelihood estimates. He observed that MLE's retain high asymptotic efficiency under modest overdispersion, provided that the logistic regression model determines the expected value of the response variable. When this is false, which is the case under certain extra-binomial specifications—for example, when $\text{logit}(Y_i) = x_i\beta + \epsilon_i$, parameter MLE's may be inconsistent. When the logistic regression model determines the expected value of the response variable, maximum likelihood parameter estimates in the presence of extra-binomial variation are asymptotically consistent but may be inefficient. In particular, McCullagh and Nelder (1989) noted that the presence of extra-binomial variation leads to underestimates of standard errors but consistent parameter estimates, so covariate effects are overestimated. They recommended using quasi-likelihood to address this problem. Following Cox's (1983) suggestion, we provide a *detailed* representation of the overdispersion by a specific model. By using finite binomial mixture models to explain extra-binomial variation, we adopt the viewpoint that there are a finite number of unobservable categories of observations that may be characterized by different values of regression coefficients. We restrict our attention to *finite* mixture models because they are physically plausible and the MLE of an arbitrary mixing distribution is a *discrete* distribution (Lindsay 1983). We believe that in many applications, the proposed models may provide interesting insights into the data and a plausible explanation for extra-binomial variation.

The article is organized as follows. Section 2 presents a motivating example that is reanalyzed using the mixed logistic regression model in Section 5. Section 3 presents and discusses the mixed logistic regression model. Parameter estimation, model selection, inference, residual analysis and goodness of fit are the topics of Section 4. Concluding remarks appear in Section 6; and computational details, a description of diagnostic measures, and Monte Carlo studies are reported in appendices.

2. AN EXAMPLE

In this article, we will apply the mixed logistic regression model to analyze data from a study in evolutionary biology reported in McCullagh and Nelder (1989, p. 143); but first we illustrate some of the limitations of standard analyses. Beetles of the genus *Tribolium* are cannibalistic in the sense that adults eat the eggs of their own species as well as those of closely related species. This study investigated whether three adult *Tribolium* species (*castaneum*, *confusum*, and *madens*) have developed an evolutionary advantage

Table 1. Number of Remaining Eggs (Out of 50) Following Two-Day Exposure to *Castaneum* Adults Classified by Replicate and Egg Species

	Egg species		
Replicate	Castaneum	Confusum	Madens
	25	24	15
1	26	14	26
	26	24	32
	29	14	32
2	28	13	19
	27	19	16
	26	10	13
3	20	7	15
	14	14	23

to recognize and avoid eggs of their own species while foraging. The experiment isolated a number of adult beetles of the same species and presented them with a vial of 150 eggs (50 of each type), the eggs being thoroughly mixed to ensure uniformity throughout the vial. Table 1 presents consumption data for adult *castaneum* species; the last three columns record number of *castaneum*, *confusum*, and *madens* eggs, respectively, that remain uneaten after two day exposure to the adult beetles. Replicates 1, 2, and 3 in the first column correspond to different occasions on which the experiment was conducted. At issue here is whether the adult species exhibit preference for eggs of the other species. Because we analyze only data from *castaneum* adults, we would expect the proportion of remaining eggs of this type to be greatest.

Table 2 reports the results of applying several logistic regression models to this data. Likelihood ratio tests suggest that egg type and replicate main effects are significant,

Table 2. Parameter Estimates for Several Logistic Regression Models

Variable	Model I estimate (SE)	Model II estimate (SE)	Model III estimate (SE)	Model IV estimate (SE)
intercept	.0534 (.1634)	.2333 (.1237)	.0272 (.0779)	−.0356 (.0943)
confusum	−.4035 (.2328)	−.7846 (.1403)	−.6475 (.1231)	−.7698 (.1389)
madens	−.1067 (.2310)	−.2746 (.1355)	—	−.2690 (.1341)
replicate 2	.1878 (.2318)	−.1380 (.1357)	—	—
replicate 3	−.4588 (.2318)	−.6746 (.1403)	−.6038 (.1228)	-
confusum.replicate 2	−.6534 (.3355)	—	—	—
confusum.replicate 3	−.5361 (.3502)	—	—	—
madens.replicate 2	−.3486 (.3277)	—	—	—
madens.replicate 3	−.1511 (.3330)	—	—	—
deviance	45.77	50.29	55.43	76.28
χ^2	45.27	49.85	54.99	75.26
d.f.	18	22	24	24
Dispersion parameter	2.52	2.27	2.29	3.14

NOTE: Model numbers designate different covariate combinations.

Table 3.   Parameter Estimates for Several Beta-Binomial Models

Variable	Model I estimate (SE)	Model II estimate (SE)	Model III estimate (SE)	Model IV estimate (SE)
intercept	.0521 (.2921)	.2284 (.2419)	.0248 (.1519)	−.0376 (.2256)
confusum	−.4025 (.4244)	−.7774 (.2718)	−.6410 (.2363)	−.7669 (.3337)
madens	−.1083 (.4279)	−.2733 (.2620)	—	−.2681 (.2911)
replicate 2	.1839 (.4210)	−.1348 (.2593)	—	—
replicate 3	−.4569 (.4173)	−.6681 (.2799)	−.5984 (.2236)	—
confusum.replicate 2	−.6348 (.6012)	—	—	—
confusum.replicate 3	−.5167 (.6407)	—	—	—
madens.replicate 2	−.3436 (.5979)	—	—	—
madens.replicate 3	−.1461 (.6043)	—	—	—
log-likelihood	−864.83	−865.76	−867.06	−871.40
ϕ	.023	.018	.022	.037

NOTE: Model numbers designate different covariate combinations.

while the interaction between these effects is insignificant. However, the full model with covariates' egg type, replicate, and interactions yields deviance and Pearson goodness-of-fit statistics 45.8 and 45.3, respectively, with 18 degrees of freedom, providing evidence of lack of fit in the logistic regression model. Furthermore, tests for extra-binomial variability (Dean 1992), yield test statistic values $N_a = 2.665$ and $N_b = N_c = 2.511$. Because these statistics are normally distributed under the null hypothesis of no extra-binomial variability, this also provides evidence of extra-binomial variability in this data. Referring to Table 1 we see that the extra-binomial variation may be attributed to excessive variability in replicate 3 and in the consumption of *madens* eggs. This may be due to sampling variation or to heterogeneity in the adults but, more importantly, it casts doubt on inferences drawn from this model.

Standard approaches for dealing with over-dispersion include quasi-likelihood (QL) and beta-binomial regression (BBR). The quasi-likelihood approach assumes the variance function $\text{var}(Y_i) = \sigma^2 m_i \pi_i (1 - \pi_i)$, while the variance function for the beta-binomial model is $\text{var}(Y_i) = m_i \pi_i (1 - \pi_i)[1 + (m_i - 1)\phi]$, where Y_i denotes the observed count at the ith observation, π_i represents the binomial probability, m_i is the binomial denominator, and ϕ is a parameter of the beta distribution. In this example, because all binomial denominators equal 50, these two variance functions are equivalent. Quasi-likelihood estimation is discussed in McCullagh and Nelder (1989) and Williams (1982), and estimation for the beta-binomial model is discussed in Kahn and Raftery (1996), Crowder (1978), Griffiths (1973), and Chatfield and Goodhardt (1970).

Under quasi-likelihood, the parameter estimates are identical to those obtained using logistic regression, but standard errors differ. To obtain the QL standard errors, estimate the over-dispersion parameter σ^2 by dividing the χ^2 goodness-of-fit statistic by the model degrees of freedom, then multiply the parameter standard errors in Table 2 by the square root of the estimate of σ^2. Estimates of σ^2 are given in the last line of Table 2. Table 3 gives parameter estimates, standard errors, and likelihood values for BBR. Using BBR there is no evidence of an interaction effect but evidence of a replicate effect. The conclusions regarding individual covariates from both QL and BBR are similar. The "t ratios" from the model with species and replicate main effects suggest that only the

confusum species effect and the replicate 3 effect are significant. A likelihood ratio test based on comparing model III to model II in Table 3 supports this observation.

Although these two models both take extra-binomial variation into account, they do so from different perspectives. The QL approach assumes a mean-variance relationship but does not correspond to an explicit probabilistic model for the data. The beta-binomial model corresponds to the same mean-variance relationship as the quasi-likelihood model, but it assumes that the π_i are sampled from a beta distribution with parameters depending on the covariates through the logit function and the additional parameter ϕ. In this example, this means that heterogeneity arises because the binomial probabilities in any cell in Table 1 are selected from a beta distribution. In this article, we propose and analyze a finite mixed logistic regression model in which the binomial probabilities in any cell are chosen from a *finite* mixture instead of from the continuous beta distribution. This model leads to a more general mean-variance relationship than the preceding two models and also provides some interesting insights into the data.

3. THE MIXED LOGISTIC REGRESSION MODEL

Let the discrete random variable Y_i denote the ith binomial response associated with the ith binomial denominator m_i, let $\{(y_i, m_i, x_i), i = 1, \ldots, n\}$ denote observations where y_i represents the observed value of Y_i, and let $x_i = (x_i^{(m)}, x_i^{(r)})'$ denote a vector of explanatory variables or covariates. The components of $x_i^{(m)}$ and $x_i^{(r)}$ are k_1-dimensional and k_2-dimensional covariate vectors corresponding to the mixing part of component binomial parameters and the regression part of the model, respectively. (Some or all components of $x_i^{(m)}$ and $x_i^{(r)}$ could be identical.) Usually, the first element of $x_i^{(m)}$ and $x_i^{(r)}$ is 1, corresponding to an intercept. The mixed logistic regression model (MLR) assumes that

(1) the unobserved mixing process can occupy any one of c states where c is finite and unknown;

(2) for each observed binomial response y_i associated with a binomial denominator m_i, there is an unobserved random variable, Π_i, representing the component that generates y_i. Furthermore, the set of "observations" (Y_i, Π_i) are pairwise and sequentially independent;

(3) Π_i follows a discrete distribution with c points of support, and $\Pr(\Pi_i = j) = p_{ij}$ where $\sum_{j=1}^{c} p_{ij} = 1$ for each i,

$$p_{ij} \equiv p_j(x_i^{(m)}, \beta) = \frac{\exp(\beta_j' x_i^{(m)})}{1 + \sum_{k=1}^{c-1} \exp(\beta_k' x_i^{(m)})} \quad \text{for } j = 1, \ldots, c-1, \qquad (3.1)$$

and

$$p_{ic} \equiv p_c(x_i^{(m)}, \beta) = 1 - \sum_{j=1}^{c-1} p_{ij}, \qquad (3.2)$$

where $\beta = (\beta_1, \ldots, \beta_{c-1})'$ and the vectors $\beta_j = (\beta_{j1}, \ldots, \beta_{jk_1})'$, for $1 \leq j \leq c-1$, are unknown parameters. Note that all components of β appear in each mixing probability p_{ij}; and

(4) conditional on $\Pi_i = j$, Y_i follows a binomial distribution with binomial denomi-
     nator m_i and probability function

$$f_j\left(y_i \mid x_i^{(r)}, m_i, \alpha_j\right) \equiv \mathrm{bi}\left(y_i \mid m_i, \pi_{ij}\right) = \binom{m_i}{y_i} \pi_{ij}^{y_i}(1 - \pi_{ij})^{m_i - y_i} \quad (3.3)$$

where the success probability of the binomial distribution and covariates are related
through the equation

$$\pi_{ij} \equiv \pi_j(x_i^{(r)}, \alpha_j) = \frac{\exp(\alpha_j' x_i^{(r)})}{1 + \exp(\alpha_j' x_i^{(r)})}, \quad \text{for } j = 1, \ldots, c, \quad (3.4)$$

$\alpha \equiv (\alpha_1, \ldots, \alpha_c)'$ and the vectors $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jk_2})'$ for $1 \leq j \leq c$ are
unknown parameters.

As a consequence of the preceding assumptions, the unconditional distribution of
y_i is a finite binomial mixture in which the mixing probabilities p_{ij} are related to the
covariates $x_i^{(m)}$ through the logistic link function, and the component distributions are
binomial distributions with the success probabilities π_{ij} defined by (3.4). If observations
can be classified into c groups corresponding to the c unobservable states, a vector of
unknown parameters α_j may be interpreted as the coefficients of the logistic regression
for group j. On the other hand, the parameters β_j may be interpreted as the coefficients
of the multinomial regression in which Π_i and $x_i^{(m)}$ are dependent and independent
variables, respectively. Note that our model allows some or all components of $x_i^{(m)}$ and
$x_i^{(r)}$ to be identical and some coefficients, α's, to be constant across components (i.e.,
$\alpha_{jl} = \alpha_l$ for $j = 1, \ldots, c$ or 0 in one or several covariates, that is, $\alpha_{jl} = 0$ for some j,
$j = 1, \ldots, c$). We let $X^{(m)} = (x_1^{(m)} \ldots x_n^{(m)})'$ and $X^{(r)} = (x_1^{(r)} \ldots x^{(r)})'$ represent two
design matrices.

Under the preceding assumptions, the probability function of Y_i satisfies

$$f(y_i \mid x_i^{(r)}, x_i^{(m)}, m_i, \alpha, \beta) = \sum_{j=1}^{c} p_{ij} \mathrm{bi}\left(y_i \mid m_i, \pi_{ij}\right), \quad (3.5)$$

where p_{ij} and $\mathrm{bi}(y_i \mid m_i, \pi_{ij})$ are specified by (3.1), (3.2), (3.3), and (3.4), respectively.
We may equivalently view the model as arising from the following sampling scheme.
Observations are independent. For observation i, component j is chosen according to
a multinomial distribution with probabilities p_{ij}. Subsequently, y_i is generated from a
binomial distribution with binomial denominator m_i and success probability π_{ij}.

A justification for the MLR model is that the coefficient vector α in the usual
logistic regression model, $\mathrm{logit}(\pi) = \alpha' x^{(r)}$, is a random variable following a discrete
distribution with c points of support, $\Pr(\alpha = \alpha_j) = p_j$ for $j = 1, \ldots, c$. By making
the further assumption that the mixture probabilities p_j are related to a covariate vector
$x^{(m)}$ through a logit link, we arrive at the model of Equation (3.5). Note that we could
have chosen other link functions in (3.4) and (3.1); *logit* link functions are used here for
convenience and tractability.

An alternative interpretation of the MLR model, which may be appropriate for the
example in the previous section, follows. The probability of a positive response varies

between experimental units according to one of c logit curves: $\text{logit}(\pi) = \alpha_j' x^{(r)}$, $(1 \leq j \leq c)$. The chance that this probability is determined by the jth curve depends on a vector of covariates $x^{(m)}$ through the function $p_j(x^{(m)}, \beta)$.

The proposed model includes several previously studied models as special cases. For instance, in the absence of covariates, (3.5) reduces to a discrete binomial mixture (e.g., Blischke 1964); if there are no covariates in the mixing probabilities and only the intercepts vary between components, it is equivalent to the semiparametric logistic regression model of Follmann and Lambert (1989); and if there are no covariates in the binomial parameter, we have a binomial variant of the mixed Poisson regression model (Wang, Cockburn, and Puterman, 1998).

For the mixed logistic regression model, the mean and variance of the response Y_i are respectively

$$E(Y_i \mid x_i) = m_i \tilde{\pi}_i \tag{3.6}$$

and

$$\text{var}(Y_i \mid x_i) = m_i \tilde{\pi}_i (1 - \tilde{\pi}_i) + \big((m_i - 1)/m_i\big) \text{var}(E(Y_i \mid \Pi_i, x_i)), \tag{3.7}$$

where $\tilde{\pi}_i = \sum_{j=1}^{c} p_{ij} \pi_{ij}$. The presence of the second term in (3.7) shows that the proposed model allows for extra-binomial variation. In a beta-binomial regression model, the variance of the response satisfies

$$\text{var}(Y_i \mid x_i) = E(Y_i \mid x_i)(1 - E(Y_i \mid x_i)/m_i)[1 + (m_i - 1)\phi], \tag{3.8}$$

where ϕ is a dispersion parameter. Rewriting (3.7) as

$$\text{var}(Y_i \mid x_i) = E(Y_i \mid x_i)(1 - E(Y_i \mid x_i)/m_i)[1 + (m_i - 1)\phi(x_i)], \tag{3.9}$$

where

$$\phi(x_i) = (\sum_{j=1}^{c} p_{ij} \pi_{ij}^2 - \tilde{\pi}_i^2)/\tilde{\pi}_i(1 - \tilde{\pi}_i)$$

is a function of model parameters, it is apparent that the beta binomial model leads to a more restrictive variance specification than the proposed model. Higher moments of these two models also differ. For the example discussed in Section 5, we substitute parameter estimates into (3.7) and (3.8) to compute the estimated conditional variance for the BBR and MLR models, respectively. An alternative model would allow ϕ in (3.8) to be an arbitrary function of the covariates, but it would lack the interpretability of the MLR model.

4. ESTIMATION AND IMPLEMENTATION

For a *fixed* number of components c, we obtain maximum likelihood estimates of the parameters using a combination of the EM algorithm (Dempster, Laird, and Rubin 1977) and the quasi-Newton algorithm (Nash 1990). As suggested by Dempster, Laird,

and Rubin, for mixture model estimation, we implement the EM algorithm by treating unobservable component membership of the observations as missing data and augment the data with indicators of component membership. When parameter estimates given by the EM algorithm are changing slowly and presumably close to the true values, we then shift to the quasi-Newton algorithm to speed up convergence. We provide details about the EM and quasi-Newton algorithms in Appendix 1 and discuss the issue of choosing the number of components in the following paragraphs.

If c is assumed known, asymptotic normality of $\hat{\alpha}$ and $\hat{\beta}$ follows from standard regularity conditions (Lehmann 1983). For fixed c we may compute parameter standard errors from the diagonal elements of the inverse of the observed information matrix.

Our approach to applying the mixed logistic regression model consists of

1. determining the number of components c using model selection criteria, and
2. carrying out inferences about model parameters when called for.

Before determining the number of components, we investigate whether the identifiability bounds in Wang (1994) place any restrictions on the model. To estimate the parameters of (3.5) and apply asymptotic theory, we require the proposed model be identifiable, that is, two sets of parameters that do not agree after permutation do not yield the same distribution. Although an unlimited class of finite binomial mixtures may not be identifiable, classes of finite mixtures of some subfamilies of binomials may be identifiable. Without covariates, Teicher (1961), Blischke (1964), and Margolin, Kim, and Risko (1989) gave necessary and sufficient conditions for identifiability of the finite binomial mixtures. With covariates, Follmann and Lambert (1991) provided sufficient conditions for the identifiability of the semi-parametric logistic regression model with common nonrandom regression coefficients and a random intercept with a finite, unknown mixing distribution. Wang (1994) defined identifiability for the MLR model and provided a loose upper bound on the number of components for which the model is identifiable. Of course, a model with more parameters than distinct observations is not identifiable.

To determine the number of components, we may appeal to subject area knowledge, use maximum likelihood theory to estimate the number of components, or appeal to model selection criteria. If the number of components is known, we would fix c at that level and then carry out inferences about parameters using likelihood ratio tests. This approach seems best suited for biological applications in which the data consists of mixtures of known numbers of different species or genders ($c = 2$), but it does not appear well suited to studies of behavior where the objective may be to gain insights into some underlying and unknown latent structure.

Using maximum likelihood to choose c has obvious limitations. When c is unknown, the likelihood ratio test statistic for inference on c does not have a χ^2 asymptotic null distribution (McLachlan and Basford 1988) so that critical values are unavailable. Hence, we might choose that value of c which maximizes the likelihood. We and other authors have noted that the likelihood increases with c and then gradually levels off. Choosing $\hat{c}$ to be the value of c at which the likelihood first reaches a plateau is ambiguous and may result in choosing models with more components then can be reasonably interpreted.

Model selection based on penalized likelihood criteria has been used in a wide range

of contexts, has good theoretical properties, and appears to work well in practice and Monte Carlo studies. The Akaike Information Criterion (AIC) (Akaike 1973, 1974) and the Bayesian Information Criterion (BIC) (Schwarz 1978) are two widely used criteria for mixture model selection. McLachlan and Basford (1988), Leroux (1989), and Leroux and Puterman (1992) applied AIC and BIC for determining the number of components in a finite mixture model without covariates. Wang, Puterman, Le and Cockburn (1996) and Wang, Cockburn and Puterman (1998) discuss the use of AIC and BIC to select a mixed Poisson regression model. Chen and Kalbfleisch (1996) proposed a promising selection criterion that penalizes low probability components in the context of models without covariates.

The AIC criterion selects a model that maximizes $\hat{l}_c(X) - M$, while the BIC criterion selects a model that maximizes $\hat{l}_c(X) - \frac{1}{2}(\log(n))M$, where $\hat{l}_c(X)$ is the log-likelihood of the c component mixture with covariates X, M is the number of estimated parameters in the model, and n is the number of observations. Using these criteria, our analysis consists of two stages. At the first stage, we determine c to maximize BIC (AIC) values for saturated mixture models that contain all covariates in both success probabilities and mixing probabilities. Note that the c values must be within the range required for identifiability. At the second stage, our analysis approach depends on our objectives. If our goal is testing hypotheses about model parameters, we use likelihood ratio tests among nested c-component mixture models. If the goal is only selecting an appropriate model to fit the data, we further choose a model to maximize BIC or AIC among c-component models. The Monte Carlo studies described in Appendix 3 suggest that when a mixed logistic regression model generates the data, the BIC criterion identifies the correct model with a high degree of precision and is more reliable than the AIC criterion. Also, when the data are generated by a mixture distribution with continuous support, both criteria choose models with more components than would be practical to interpret.

After fitting a mixed logistic regression model to data, it is essential to have diagnostic measures that assess the quality of the fit. For this purpose, Wang (1994) defined Pearson, deviance, and deleted likelihood residuals for the model and discussed the use of these residuals to identify poorly fitted observations. Wang also introduced the average relative coefficient change to identify influential observations and for the mixed logistic regression models. We calculate this quantity by deleting each observation from the dataset and re-estimating model parameters. The parameter values from the complete dataset provide good starting values for the estimation algorithms. We describe these diagnostic measures in detail in Appendix 2.

In addition, we use the Pearson goodness-of-fit statistic (the sum of squared Pearson residuals) to evaluate the extent to which a mixed logistic regression model adequately fits the entire set of data. Under the null hypothesis that the model is appropriate, this statistic is asymptotically distributed χ^2_{n-p}, where n is the number of observations and p is the number of unknown parameters in the model.

5. APPLICATION OF MLR TO THE EXAMPLE

We now return to the analysis of the example in Section 2. Because relative preference for eggs of other species may vary between beetles, regression coefficients for the

relative egg species effects may vary in order to capture this heterogeneity. Our mixture model proposes several discrete levels for these effects. Further, to account for variability between replicates, we allow the expected proportions of beetles with different consumption preferences to vary with replicate. We model this by including species-dependent covariates $x_{i1}^{(r)}$ and $x_{i2}^{(r)}$ in the binomial parameter and replicate-dependent covariates $x_{i1}^{(m)}$ and $x_{i2}^{(m)}$ in the mixing probabilities. Formally,

$$x_{i1}^{(r)} = \begin{cases} 1 & \text{if eggs in observation } i \text{ are from the } \textit{confusum} \text{ species,} \\ 0 & \text{otherwise;} \end{cases}$$

$$x_{i2}^{(r)} = \begin{cases} 1 & \text{if eggs in observation } i \text{ are from the } \textit{madens} \text{ species,} \\ 0 & \text{otherwise;} \end{cases}$$

$$x_{i1}^{(m)} = \begin{cases} 1 & \text{if observation } i \text{ is from replicate 2,} \\ 0 & \text{otherwise;} \end{cases}$$

$$x_{i2}^{(m)} = \begin{cases} 1 & \text{if observation } i \text{ is from replicate 3,} \\ 0 & \text{otherwise.} \end{cases}$$

Both vectors $x_i^{(r)}$ and $x_i^{(m)}$ contain a 1, corresponding to the constant as the first component. The identifiability bound in Wang (1994) equals 25, so it imposes no practical restrictions in this model. Because there are 27 observations and a 4-component model has 21 parameters, a model with more than 4 components would not be identifiable.

5.1 RESULTS

Table 4 gives parameter estimates, the log-likelihood, the AIC, and the BIC values for saturated 1- and 3-component models and a wide range of 2-component models. To select the number of components, we compare the values of AIC and BIC among the three saturated models. Both AIC and BIC lead to the choice of a 2-component mixed logistic regression model. Further, within the class of 2-component models, both criteria choose a model with a replicate 3 effect in the mixing probabilities and a common *confusum* species effect in the binomial parameter. Parameter estimates and standard errors for this model are given in the last column of Table 5. Subject matter knowledge might shed some light on the interpretation of these estimates but, unfortunately, is unavailable to us.

Alternatively, we may carry out hypothesis tests within the class of 2-component models by using the χ^2 approximation to the likelihood ratio test statistic that is valid when the nested models have the same number of components. From this analysis, there is evidence of a replicate effect in the mixing probabilities as the likelihood ratio statistic equals $2(883.32 - 878.88) = 8.88$ with p value of .0117, and further that the replicate 1 and 2 effects do not differ, because the likelihood ratio statistic equals $2(879.29 - 878.88) = .82$ with p value of .3652. Within 2-component models that include a replicate 3 effect in the mixing probabilities, there is evidence of a species effect in the binomial parameters because the likelihood ratio statistic equals $2(885.99 - 879.29) = 13.4$ with p value of .0095. Further, there is insufficient evidence to conclude that preference differs between the *castaneum* and *madens* egg types, and it appears that the preference

Table 4. Mixed Logistic Regression Model Parameter Estimates

Component number (j)	Mixing probabilities			Binomial parameters			log-likelihood*	AIC	BIC
	β_{j0}	β_{j1}	β_{j2}	α_{j0}	α_{j1}	α_{j2}			
			Logistic regression model (1-component mixture):						
1	—	—	—	−.0356	−.7698	−.2690	−896.80	−899.80	−901.74
			2-component mixture:						
1	−.0878	—	—	−.9326	—	—	−887.90	−890.90	−892.84
2	—	—	—	.0933	—	—			
1	−.6528	—	1.8507	−.9211	—	—	−885.99	−889.99	−892.58
2	—	—	—	.0962	—	—			
1	−.1348	—	—	−.7387	−.4175	—	−884.41	−889.41	−892.65
2	—	—	—	.1699	−.4694	—			
1	.5529	—	—	−.1389	−.9896	−.5779	−883.32	−890.32	−894.86
2	—	—	—	.1242	−.3492	.2051			
1	.6489	−.3986	—	−.1633	−.9732	−.5479	−883.26	−891.26	−896.44
2	—	—	—	.1428	−.3791	.1924			
1	−1.0718	—	3.0492	−.6664	−.5400	—	−881.09	−887.09	−890.98
2	—	—	—	.1974	−.6318	—			
1	−.9517	—	3.0545	−.6458	−.5602	—	−881.13	−886.13	−889.98
2	—	—	—	.1942	—	—			
1	−.5028	—	17.01	−.2639	−.8954	−.4102	−879.29	−897.29	−892.47
2	—	—	—	.1573	−.4551	.2470			
1	−1.1287	1.1055	17.36	−.2672	−.8887	−.4088	−878.88	−887.88	−893.71
2	—	—	—	.1482	−.4469	.2561			
			3-component mixture:						
1	3.5145	−4.2349	17.28	−.3484	−.7981	−.3402			
2	4.5742	−22.38	−3.0996	.0791	−.1888	.2700	−876.82	−891.82	−901.53
3	—	—	—	.2438	−1.0119	.3901			

NOTE: Within 2-component mixtures, each pair of lines corresponds to a particular choice of covariates.
* Log likelihood excludes the constant term.

for *confusum* eggs is identical in the two components. Hence, formal inferences lead to the same conclusions as model selection.

The model suggests that number of remaining eggs is generated by two underlying binomial distributions with conditional binomial parameters

$$\pi_{i1} = \frac{\exp(-.6458 - .5602x_{i1}^{(r)})}{1 + \exp(-.6458 - .5602x_{i1}^{(r)})}$$

and

$$\pi_{i2} = \frac{\exp(.1942 - .5602x_{i1}^{(r)})}{1 + \exp(.1942 - .5602x_{i1}^{(r)})}.$$

In addition, the mixing probabilities satisfy

$$p_{i1} = \frac{\exp(-.9517 + 3.0492x_{i2}^{(m)})}{1 + \exp(-.9517 + 3.0492x_{i2}^{(m)})}$$

Table 5. Parameter Estimates (Standard Errors) Obtained Using Different Models

	Covariate	Logistic regression	Quasi-likelihood	Beta-binomial regression	Mixed logistic regression comp. 1	comp. 2
bin. par.	intercept	.0272 (.0779)	.0272 (.1179)	.0248 (.1519)	−.6458 (.1312)	.1942 (.1012)
	confusum	−.6475 (.1231)	−.6475 (.1863)	−.6410 (.2363)	−.5602 (.2296)	
	replicate 3	−.6038 (.1228)	−.6038 (.1858)	−.5984 (.2236)	NA	
mix. prob.	intercept	NA	NA	NA	−.9517 (.7792)	
	replicate 3	NA	NA	NA	3.5045 (1.7929)	
	Dispersion parameter	$\sigma^2 = 1.0$	$\sigma^2 = 2.29$	$\phi = .022$	NA	

and

$$p_{i2} = \frac{1}{1 + \exp(-.9517 + 3.0492 x_{i2}^{(m)})}.$$

According to this model, the *castaneum* adults have a preference for eggs of the *confusum* species, but are indifferent between the eggs of the *castaneum* and *madens* species. Moreover, *castaneum* adults may be classified into "high" and "low" consumption groups on the basis of the posterior probabilities of component membership obtained using the EM algorithm. Figure 1 illustrates this classification. Observations marked as "1" correspond to the "high" consumption group and those denoted by "2" correspond to the "low" consumption group. Note that only one observation is classified in the low consumption group for the third replicate. For the high consumption group, 23% of the *confusum* eggs and 34% of the *castaneum* and *madens* eggs remain; and in the low consumption group, 40% of the *confusum* eggs and 54% of the *castaneum* and *madens* eggs remain. Furthermore, the expected proportion of beetles allocated to the high and low consumption groups differs over the three replicates: for replicates 1 and 2, 28% may be classified in the high consumption group, with this proportion increasing to 89% for replicate 3.

Note that the fitted model is not in the class of models considered by Follmann and Lambert (1989). Their model applied here would allow only discrete mixtures of intercepts with mixing probabilities constant across replicates, while the model obtained here has replicate dependent mixture probabilities. Aitkin (1995) suggested an extension of the Follmann and Lambert model to random coefficient regression models that does not include covariates in the mixing probabilities. These models have been implemented in a set of GLIM macros by Aitkin and Francis (1995). An anonymous reviewer reported that the application of these macros to the fourth 2-component model in Table 4, "gives an identical value of the log-likelihood and very similar parameter values."

The fitted model has a Pearson goodness-of-fit test statistic of 24.2 with 22 degrees of freedom, providing no evidence of lack of fit. Figure 2 provides index plots of Pearson, deviance, and likelihood residuals from the fitted model. These plots show that the three
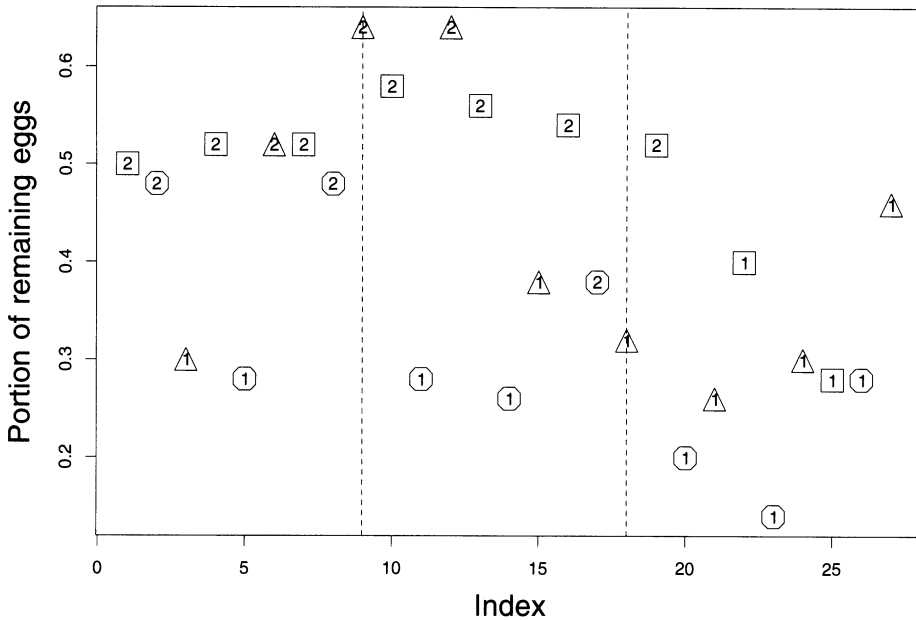
Figure 1. *Classification for Egg Data.* □ = *castaneum eggs,* ○ = *confusium eggs, and* △ = *madens eggs.*

types of residuals are similar, and provide weak evidence that the 19th observation (the observation corresponding to replicate 3, *castaneum* species with 26 eggs remaining) is an outlier. The average relative coefficient changes, w_i, which are displayed in Figure 2(d), provide convincing evidence that this observation has a large effect on coefficient estimates. On omitting this observation, the average relative coefficient change for each parameter estimate is 337% and the deviance decreases by $r_{L19}^2 = (2.3224)^2 = 5.3933$.

When observation 19 is deleted, parameter estimates become

$$\hat{\beta}_1 = (.8815, -17.01),$$
$$\hat{\alpha}_1 = (-.2135, -.5410),$$
$$\hat{\alpha}_2 = (-.6561, -.5410).$$

Note that changes for mixing probability parameter estimates and regression intercept estimates are large, while changes in the species effect parameter are small. Hence, inferences regarding the species effect are unaffected. On omitting this observation, the mixing probability of the first component is almost 1 for replicate 3, suggesting that observation 19 may be the cause of overdispersion for the data of replicate 3. On the other hand, the data for the first two replicates are still overdispersed and are fitted by the 2-component MLR model with a common coefficient for *confusum* eggs. Unfortunately, we do not have further information that would enable us to determine if observation 19 should be deleted from the analysis.
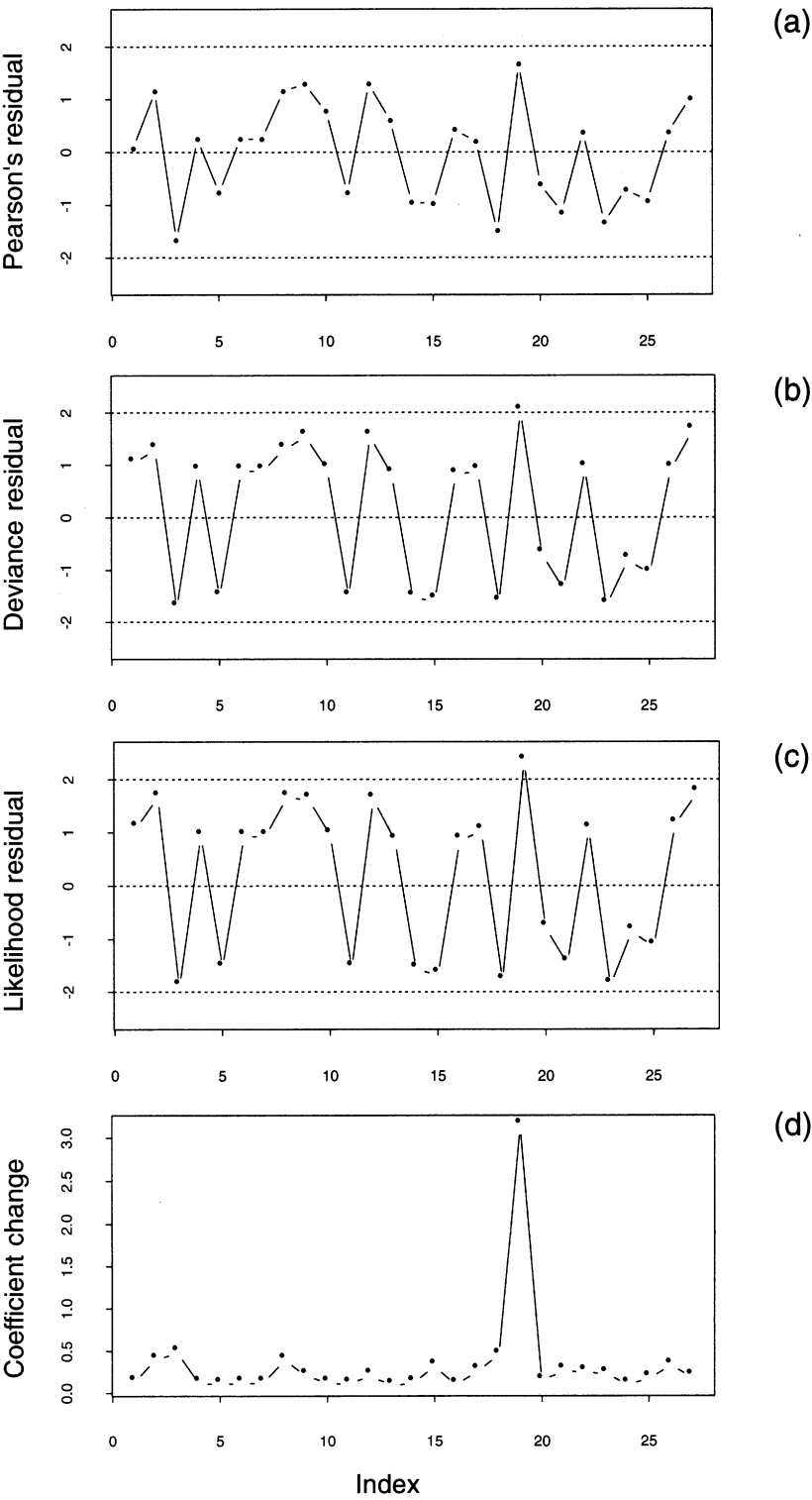
Figure 2.   Residual and Coefficient Change Plots.

Table 6. Estimated Mean and (Variance) for Different Models and Different Egg Species

Model	First two replicates		Third replicate	
	castaneum & madens	confusum	castaneum & madens	confusum
Logistic	25.3 (12.5)	17.5 (11.4)	18.0 (11.5)	11.4 (8.8)
QL	25.3 (28.6)	17.5 (26.0)	18.0 (26.3)	11.4 (20.1)
BBR	25.3 (26.0)	17.5 (23.7)	18.0 (24.0)	11.4 (18.3)
MLR	24.2 (32.2)	17.6 (25.7)	17.9 (18.8)	12.1 (14.4)
Data	25.1 (31.5)	18.0 (26.0)	18.5 (28.3)	10.3 (12.3)

NOTE: The last row gives corresponding data values.

5.2   COMPARISON TO OTHER MODELS

Table 5 summarizes parameter estimates and standard errors for the four estimation methods applied to this data. Note that in terms of t statistics (parameter estimate/standard error), from Model II in Table 2 and Table 3, we conclude that the *madens* and the second replicate effects are not siginicant at the .05 level using quasi-likelihood and beta-binomial regression, and with logistic regression the *madens* effect is significant but the second replicate effect is not. Logistic regression, QL, and BBR yield very similar parameter estimates but different standard errors. The logistic regression has the smallest standard errors because it takes no account of any overdispersion, while the QL has smaller standard errors than the BBR. The estimate of the *confusum* effect parameter is smaller under the mixed logistic regression model because this model takes into account extra variability through the replicate dependent mixing probabilities. Standard errors of the mixed logistic regression model are similar to that of the beta-binomial regression model.

It is reasonable to inquire whether the different models yield different fits to the data. We gain insight into this issue in two ways. First, we compare the fitted cumulative probability distributions of remaining egg counts for the first two replicates. Figure 3 shows the fitted cumulative probability distributions for the logistic, beta-binomial, and mixed logistic regression models. Fits for quasi-likelihood are the same as logistic regression. The fitted cumulative probability distributions are quite different, especially in the range from 15 to 23 for *castaneum* and *madens* species and the range from 9 to 15 for *confusum* species. For instance, for *castaneum* and *madens*, the probability of 20 and fewer remaining eggs is .05 for logistic regression, .13 for beta-binomial, and .23 for the mixed logistic regression model; and for *confusum* eggs, the probability of 12 and fewer remaining is .03 for logistic regression, .11 for beta-binomial, and .15 for the mixed model. The mixed model appears bimodal while the other distributions appear approximately normal.

Table 6 summarizes the estimated means and variances of remaining egg counts for each of the three models and the actual values. The means and variances for the logistic model are computed in each cell using the binomial formulae $\mathrm{E}(Y_i) = 50\pi_i$ and $\mathrm{var}(Y_i) = 50\pi_i(1 - \pi_i)$, where π_i denotes the cell probability. The QL variance is obtained by multiplying the binomial variance by the estimate of the dispersion
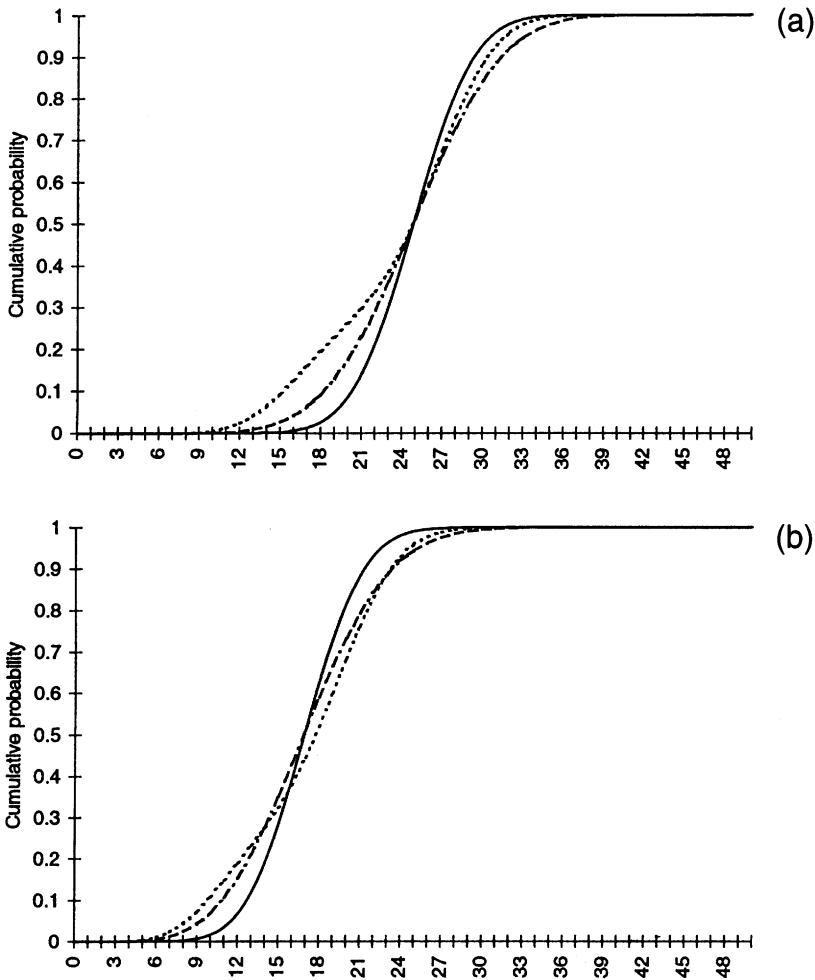
Figure 3. Comparisons of cumulative probabilities for Logistic (——), Beta-Binomial (– - – -), and Mixed Logistic (- - - -) Regression for (a) casteneum and madens and (b) confusium, based on models in Table 5. Replicate 3 covariate value set equal to 0 so that probabilities are averages over replicates 1 and 2.

parameter σ^2 in Table 5. The variances of the BBR and MLR models are computed using expressions (3.7) and (3.8). Observe that although the estimated means are quite similar, the estimated variances and mean-variance ratios for the three models differ greatly. For the first two replicates, the estimated variance and mean-variance ratio from the MLR are the closest to those in the data within the *castaneum* and *madens* egg categories, while the estimated variance and mean-variance ratios from the QL and the MLR are similar and closer to those in the data within the *confusum* egg category than that from the BBR. For the third replicate, the estimated variance and mean-variance ratio from the MLR are the closest to those in the data within the *confusum* egg category, while the mean and mean-variance ratio from the QL are the closest to those in the data under *castaneum* and *madens* egg categories. However, on omitting observation 19, which was previously identified as an outlier, the mean and variance in the data in

the *castaneum* and *madens* category become 17.0 and 18.5, respectively. These are very similar to those from the MLR, suggesting that the mean-variance relation of the MLR may capture the true one without the outlier. Although the difference in the fit may be partially attributed to the fact that the MLR has one more parameter than the BBR and QL models, this analysis may also suggest that the MLR model provides a better fit to the data.

As a final comparison, we also applied a 2-component MLR model with replicate and species covariates in the regression specification and constant mixture probabilities. We found that likelihood and model selection criteria led to conflicting conclusions, that the BIC and AIC values were close to those of the model chosen previously, and that the fitted model was difficult to interpret. We prefer the model with species covariates in the regression and replicate effects in the mixing probabilities because they allow the mix of beetles with different consumption patterns to vary with replicate.

6. CONCLUSION

This article presents a class of finite mixed logistic regression model which allow for covariates in both the success probabilities of the component distributions and mixing probabilities. This model can be used to model and lend insight into the sources of extra-binomial variation in logistic regression models and provides an interpretable alternative to quasi-likelihood and beta-binomial models.

Some issues raised by this article include the need for likelihood based methods for inference regarding the number of components, a thorough investigation of the robustness of this model to mixing distribution misspecification, the analysis of a BBR model in which ϕ in (3.8) is a function of covariates, and significant original applications of these methods.

6.1 CODES

FORTRAN codes for the mixed logistic models may be downloaded from
<http://markov.commerce.ubc.ca/marty/research.html>.

APPENDIX A: DETAILS OF THE ESTIMATION ALGORITHM

Suppose that $(Y, M, X^{(m)}, X^{(r)}) \equiv \{(y_i, m_i, x_i^{(m)}, x_i^{(r)}); i = 1, \ldots, n\}$ is the observed data. Let

$$(Y, Z, M, X^{(m)}, X^{(r)}) \equiv \{(y_i, z_i, m_i, x_i^{(m)}, x_i^{(r)}); i = 1, \ldots, n\}$$

denote the complete data for the model, where the unobserved quantity $z_i = (z_{i1}, \ldots, z_{ic})'$ satisfies

$$z_{ij} = \begin{cases} 1 & \text{if } \Pi_i = j \\ 0 & \text{otherwise.} \end{cases}$$

The log-likelihood of the complete data is

$$
\begin{aligned}
l^c \quad \equiv \quad l(\alpha, \beta \mid Y, Z, M, X^{(m)}, X^{(r)}) &= \sum_{i=1}^{n} \sum_{j=1}^{c} z_{ij} \log(p_{ij}) \\
&+ \sum_{i=1}^{n} \sum_{j=1}^{c} z_{ij} \log\big(\mathrm{bi}(y_i \mid m_i, \pi_{ij})\big),
\end{aligned}
$$

where p_{ij} and $\mathrm{bi}\big(y \mid m_i, \pi_{ij}\big)$ are defined by (3.1), (3.2), and (3.3), respectively.

The EM approach finds the maximum likelihood estimates using an iterative procedure consisting of two steps: an E-step and an M-step. At the E-step, it replaces the missing data by its expectation, conditional on the observed data and the initial values of parameters. At the M-step, it finds the parameter estimates that maximize the expected log likelihood for the complete data, conditional on the expected values of the missing data. Iteration stops when the log likelihood for the observed data does not increase significantly. In our case this procedure can be stated as follows.

E-step: Given the values $\alpha^{(0)}$ and $\beta^{(0)}$, replace the missing data, Z, by its expectation conditioned on these initial values of the parameters and the observed data, $(Y, M, X^{(m)}, X^{(r)})$. In this case, the conditional expectation of the jth component of z_i equals the probability that the observation y_i was generated by the jth component of the mixture distribution, conditional on the parameters, the data, and the covariates. Denote the conditional expectation of the jth component of z_i by $\tilde{z}_{ij}(\alpha^{(0)}, \beta^{(0)})$. Then

$$
\begin{aligned}
\tilde{z}_{ij} &= \mathrm{E}\left(z_{ij} \mid Y, M, X^{(m)}, X^{(r)}, \alpha^{(0)}, \beta^{(0)}\right) \\
&= \mathrm{Pr}\left(z_{ij} = 1 \mid Y, X^{(m)}, X^{(r)}, M, \alpha^{(0)}, \beta^{(0)}\right) \\
&= \frac{p_j(x_i^{(m)}, \beta^{(0)})\mathrm{bi}(y_i \mid m_i, \pi_j(x_i^{(r)}, \alpha_j^{(0)}))}{\sum_{l=1}^{c} p_l(x_i^{(m)}, \beta^{(0)})\mathrm{bi}(y_i \mid m_i, \pi_l(x_i^{(r)}, \alpha_l^{(0)}))}, \quad \text{for } j = 1, \dots, c, \quad \text{(A.1)}
\end{aligned}
$$

where $p_j(x_i^{(m)}, \beta^{(0)})$ and $\mathrm{bi}(y_i \mid m_i, \pi_j(x_i^{(r)}, \alpha_j^{(0)}))$ are defined by (3.1), (3.2), and (3.3), respectively.

M-step: Given conditional probabilities $\{\tilde{z}_i(\alpha^{(0)}, \beta^{(0)}) = (z_{i1}, \dots, z_{ic})'; i = 1, \dots, n\}$, obtain estimates of the parameters by maximizing, with respect to α and β,

$$
\begin{aligned}
Q(\alpha, \beta \mid \alpha^{(0)}, \beta^{(0)}) &= \mathrm{E}\left(l^c \mid Y, X^{(m)}, X^{(r)}, M, \alpha^{(0)}, \beta^{(0)}\right) \\
&\equiv Q_1(\beta \mid \beta^{(0)}) + Q_2(\alpha \mid \alpha^{(0)}),
\end{aligned}
$$

where

$$
Q_1(\beta \mid \beta^{(0)}) = \sum_{i=1}^{n} \sum_{j=1}^{c} \tilde{z}_{ij} \log(p_{ij})
$$

and

$$
Q_2(\alpha \mid \alpha^{(0)}) = \sum_{i=1}^{n} \sum_{j=1}^{c} \tilde{z}_{ij} \log\big(\mathrm{bi}(y_i \mid m_i, \pi_{ij})\big).
$$

The estimated parameters, $\hat{\alpha}$ and $\hat{\beta}$, satisfy the following M-step equations:

$$\frac{\partial Q_1}{\partial \alpha}\big|_{\hat{\alpha}} = 0 \tag{A.2}$$

and

$$\frac{\partial Q_2}{\partial \beta}\big|_{\hat{\beta}} = 0. \tag{A.3}$$

Because closed form solutions of these equations are unavailable, we use an implementation of the quasi-Newton approach provided by Nash (1990) to obtain estimates. We note that in contrast to iteratively reweighted least squares, in the quasi-Newton algorithm the step size at each iteration is scaled to ensure an increase in the likelihood.

Dempster, Laird and Rubin (1977) and Wu (1983) discussed the convergence of the EM algorithm in a general setting. Applying Wu's results in our case lets us conclude that the sequence of the parameter estimates given by the EM algorithm converges to a local maximum or saddle point of the observed log likelihood function, $l(\alpha, \beta \mid Y, M, X^{(m)}, X^{(r)})$, regardless of starting values. In practice, however, we need to choose initial values carefully in order to increase the chance that the algorithm converges to the *global* maximum. Wang (1994) proposed an approach of choosing initial values for this purpose.

On the other hand, the EM algorithm converges linearly (Redner and Walker 1984; McCullagh and Nelder 1989), while a quasi-Newton algorithm that maximizes the observed log likelihood function has a higher convergence rate, though it requires good starting values. Using the complementary strengths of both algorithms, we first use the EM algorithm until either the likelihood value does not increase significantly in terms of a given tolerance ϵ_2 or the parameter estimates do not change significantly in terms of a given tolerance ϵ_1, and then shift to a quasi-Newton algorithm that maximizes the observed likelihood function. In doing so, we can obtain approximate standard error of the estimates as a by-product of the quasi-Newton approach. Note that in some cases the quasi-Newton approximate standard errors may not be accurate. Hence, we recommend calculating the information matrix by using a closed formula whenever possible. We implement the E and M steps in the following way to obtain parameter estimates.

Step 0: Specify starting values $\alpha^{(0)} = (\alpha_1^{(0)}, \ldots, \alpha_c^{(0)})$ and $\beta^{(0)} = (\beta_1^{(0)}, \ldots, \beta_{c-1}^{(0)})$ and two tolerances, ϵ_1 and ϵ_2.

Step 1: (E-step) Compute $\tilde{z}_i = (\tilde{z}_{i1}, \ldots, \tilde{z}_{i1})'$, $(1 \leq i \leq n)$, using (A.1). To avoid an overflow problem in the calculation of $\tilde{z}_{ij}$, divide both the numerator and denominator in (A.1) by the largest term in the sum in the denominator.

Step 2: (M-step) Find values of $\hat{\alpha}$ and $\hat{\beta}$ to solve (A.2) and (A.3), respectively, using the quasi-Newton algorithm (Nash 1990).

Step 3: (a) If at least one of the following conditions is true, set $\alpha^{(0)} = \hat{\alpha}$ and $\beta^{(0)} = \hat{\beta}$, and go to Step 1; Otherwise, go to (b).

    (1) $\|\hat{\alpha} - \alpha^{(0)}\| \equiv \sum_{j=1}^{c} \sum_{l=1}^{k_1} |\hat{\alpha}_{j,l} - \alpha_{j,l}^{(0)}| \geq \epsilon_1$;

    (2) $\|\hat{\beta} - \beta^{(0)}\| \equiv \sum_{j=1}^{c} \sum_{l=1}^{k_2} |\hat{\beta}_{j,l} - \beta_{j,l}^{(0)}| \geq \epsilon_1$;

    (3) $|l(\hat{\alpha}, \hat{\beta} \mid Y, M, X^{(m)}, X^{(r)}) - l(\alpha^{(0)}, \beta^{(0)} \mid Y, M, X^{(m)}, X^{(r)})| \geq \epsilon_2$.

(b) Maximize the observed log likelihood function $l(\alpha, \beta \mid Y, M, X^{(m)}, X^{(r)})$ using the quasi-Newton algorithm (Nash, 1990) with $\hat{\alpha}$ and $\hat{\beta}$ as initial values.

Then stop.

APPENDIX B: DIAGNOSTIC MEASURES

The *Pearson residuals* satisfy

$$r_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}, \tag{B.1}$$

where

$$\hat{\mu}_i = m_i \sum_{j=1}^{c} \hat{p}_{ij}\hat{\pi}_{ij}, \tag{B.2}$$

$$\hat{\pi}_{ij} = \exp(\hat{\alpha}_j' x_i^{(r)})/(1 + \exp(\hat{\alpha}_j' x_i^{(r)})),$$

$$\hat{p}_{ij} = \frac{\exp(\hat{\beta}_j' x_i^{(m)})}{\sum_{k}^{c-1} \exp(\hat{\beta}_k' x_i^{(m)}) + 1} \quad \text{for } j = 1, \ldots, c-1,$$

$$\hat{p}_{ic} = \frac{1}{\sum_{k}^{c-1} \exp(\hat{\beta}_k' x_i^{(m)}) + 1},$$

and

$$V(\hat{\mu}_i) = m_i \sum_{j=1}^{c} \hat{p}_{ij}\hat{\pi}_{ij}(1 - \sum_{j=1}^{c} \hat{p}_{ij}\hat{\pi}_{ij}) + m_i(m_i - 1)\left\{\sum_{j=1}^{c} \hat{p}_{ij}\hat{\pi}_{ij}^2 - [\sum_{j=1}^{c} \hat{p}_{ij}\hat{\pi}_{ij}]^2\right\}. \tag{B.3}$$

The *deviance residual* is defined as

$$\begin{aligned} r_{Di} &= \text{sign}(y_i - \hat{\mu}_i)\sqrt{2[l(y_i, y_i) - l(\hat{\mu}_i, y_i)]} \\ &= \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \end{aligned} \tag{B.4}$$

where $l(\mu_i, y_i)$ is the log likelihood function of mixed logistic regression model for observation y_i and $d_i = 2(l(y_i, y_i) - l(\hat{\mu}_i, y_i))$ is the contribution to the deviance goodness-of-fit statistic D, which is defined as

$$D = 2\sum_{i=1}^{n}[l(y_i, y_i) - l(\hat{\mu}_i, y_i)]. \tag{B.5}$$

Note that $l(y_i, y_i)$ is the same for both the usual logistic regression and mixed logistic regression models because

$$f\left(y_i \mid x_i^{(r)}, x_i^{(m)}, m_i, \alpha, \beta\right) = \sum_{j=1}^{c} p_{ij}\text{bi}\left(y_i \mid m_i, \pi_{ij}\right) \tag{B.6}$$

$$\leq \sum_{j=1}^{c} p_{ij}\text{bi}\left(y_i \mid m_i, y_i\right) \tag{B.7}$$

$$= \text{bi}\left(y_i \mid m_i, y_i\right). \tag{B.8}$$

This indicates that the baseline is the same for the usual logistic regression models and mixed logistic regression models.

The *deleted likelihood residual* is derived by comparing the deviance obtained on fitting a mixed logistic regression model to the complete set of n cases with the deviance obtained when the same model is fitted to the $n - 1$ cases, excluding the ith, for $i = 1, \ldots, n$. This gives rise to a quantity that measures the change in the deviance when each case in turn is excluded from the dataset. The value of the likelihood residual for the ith case is defined as

$$r_{Li} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{D - D^{(i)}}, \tag{B.9}$$

where $\hat{\mu}_i$ is defined by (B.2); and D and $D^{(i)}$ are the deviances based on n and $n - 1$ cases, respectively.

Note that for large binomial denominators m_i, all three residuals approximately follow the standard normal distribution when the fitted model is adequate. The numerical results from the Monte Carlo study (Wang 1994) show that the Pearson residuals may depart further from normality than the other residuals.

To examine how the ith observation affects the set of parameter estimates, we define the following quantity:

$$
\begin{aligned}
w_i &= \frac{1}{M}\left\{\| (\hat{\alpha} - \hat{\alpha}^{(i)})/\text{se}(\hat{\alpha}) \| + \| (\hat{\beta} - \hat{\beta}^{(i)})/\text{se}(\hat{\beta}) \|\right\} \\
&= \frac{1}{M}\left\{\sum_{j=1}^{c}\sum_{l=1}^{k_1}\frac{|\hat{\alpha}_{j,l} - \hat{\alpha}_{j,l}^{(i)}|}{\text{se}(\hat{\alpha}_{j,l})} + \sum_{j=1}^{c-1}\sum_{l=1}^{k_2}\frac{|\hat{\beta}_{j,l} - \hat{\beta}_{j,l}^{(i)}|}{\text{se}(\hat{\beta}_{j,l})}\right\},
\end{aligned}\tag{B.10}
$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood parameter estimates of the mixed logistic regression model based on the complete dataset of n cases and $\hat{\alpha}^{(i)}$ and $\hat{\beta}^{(i)}$ on the dataset of $n - 1$ cases excluding the i case; $\text{se}(\hat{\alpha})$ and $\text{se}(\hat{\beta})$ are the estimated standard errors of the corresponding estimates based on the n cases; and M is the number of estimated parameters $(c * k_1 + (c - 1) * k_2)$. Because each term in (B.10) measures a relative change in individual coefficient, w_i can be interpreted as an average relative coefficient change for a set of estimates. This is a useful quantity for assessing the extent to which the set of parameter estimates is affected by the exclusion of the ith observation. Relatively large values of this quantity will indicate that the corresponding observations are influential and causing instability in the fitted model. An index plot of w_i is the most useful way of presenting these values. Subsequent analysis may investigate the effect on individual components.

APPENDIX C: MONTE CARLO STUDIES

In this section we describe two Monte Carlo studies: one which investigated the variability of estimates, the reliability of our algorithms, and the behavior of the model selection procedures, and the other which examined the effect of misspecification of the mixing distributions on parameter estimates and model selection.

Table C.1.   The Results of Monte Carlo Study 1

The First Model

	Success probabilities						Mixing probabilities					
Comp. (j)	α_{j0}	Mean	Std. Dev.	α_{j1}	Mean	Std. Dev.	β_{j0}	Mean	Std. Dev.			
1	−1.2962	−1.2402	.1515	−.4505	−.4855	.1244	−.9163	−.9643	.1248			
2	−1.3148	−1.3619	.1066	1.0811	1.1211	.0700	−.5108	−.5380	.1376			
3	.69733	.6892	.0787	.7499	.7683	.0823	—	—	—			

The Second Model

	Success probabilities						Mixing probabilities					
Comp. (j)	α_{j0}	Mean	Std. Dev.				β_{j0}	Mean	Std. Dev.	β_{j0}	Mean	Std. Dev.
1	−2.1972	−2.2061	.0835	—	—	—	−2.1129	−2.3068	.3959	1.6057	1.7467	.4215
2	−.8474	−.8473	.0438	—	—	—	−.9692	−1.0958	.2322	1.3805	1.4941	.2032
3	1.3863	1.3892	.0345	—	—	—	—	—	—	—	—	—

The Third Model

	Success probabilities						Mixing probabilities					
Comp. (j)	α_{j0}	Mean	Std. Dev.	α_{j1}	Mean	Std. Dev.	β_{j0}	Mean	Std. Dev.	β_{j1}	Mean	Std. Dev.
1	−1.2962	−1.2114	.2073	−.4505	−.5067	.1404	−2.1129	−2.2618	.4389	1.6057	1.7061	.3395
2	−1.3148	−1.3164	.0835	1.0811	1.0718	.0613	−.9692	−1.1588	.2148	1.3805	1.5383	.1050
3	.6973	.6881	.0826	.7499	.7604	.1073	—	—	—	—	—	—

NOTE: Entries are the true parameter values and the mean and standard deviations over 101 replicates.

C.1   Monte Carlo Study 1

We summarize the structure and findings of a Monte Carlo study reported in Wang (1994). Data was generated from three different 3-component mixed logistic regression models in which $m_i \equiv 30$: one with covariates in the success probabilities only, one with covariates in the mixing probabilities only, and one with covariates in each. In each case there were 101 replicates, and each replicate corresponded to 100 observations. In each model, covariate values consisted of 10 observations at .2, 10 observations at .4, and so forth. The first model had mixing probabilities .2, .3, and .5 and success probability functions logit$(-1.2962 - .4505x_i)$, logit$(-1.3148 + 1.0811x_i)$, and logit$(.6973 + .7499x_i)$. The second model had mixing probability parameter vectors $(-2.1129, 1.6057)$ and $(-.9692, 1.3805)$ and success probabilities .1, .3, and .8. The third model combines the success probability parameter vectors from the first model with the mixing probability parameter vectors from the second model. Parameter values were chosen so that the corresponding functions would be well separated.

Table C.1 reports the true parameter values and the mean and standard deviations of parameter estimates from the simulations. Observe that in all models, the mean of the parameter estimates is close to the true parameter value, with the success probability parameters having smaller bias than the mixing probability parameters. Also, the standard deviations of the success probability parameters were smaller than that of those of the mixing probability parameters especially when there were covariates in the mixing probabilities.

Table C.2.   Estimates of α_1 based on 500 Replicates in Monte Carlo Study 2

		True Model:	Mixed Logistic	Regression	
Fitted Model	Average	Median	Std. Dev.	Interquartile	Range
Mixed Logistic	−.5982	−.5944	.1332	(−.6796, −.5109)	(−.9444,.0957)
Beta-Binomial	−.5276	−.7990	.3310	(−1.1217, −.4565)	(−1.9222, .6596)

		True Model:	Beta-binomial	model	
Fitted Model	Average	Median	Std. Dev.	Interquartile	Range
Mixed Logistic	−.6658	−.6089	.5085	(−.9487,−.3474)	(−2.7798, 1.0902)
Beta-binomial	−.6136	−.6289	.3912	(−.8929, −.4013)	(−2.1028, .3565)

NOTE: True value of $\alpha_1 = -.6$.

We also used this data to compare the performance of the AIC and BIC model selection criteria. For the first two models, we compared the fit of 2- to 4-component models with and without the respective covariates for each replicate. For each replicate of data generated from the third model, we compared the fit of 2 to 4 component models without covariates, with covariates in the binomial parameter only, with covariates in the mixing probabilities only, and with covariates in both the mixing probabilities and binomial parameters.

We found that AIC chose the correct model in 95 of the 101 replicates for the first model, 83 of the 101 replicates for the second model, and 94 of the 101 replicates for the third model. When the AIC failed to select the correct model, it chose a model with too many parameters, suggesting that AIC may under-penalize the number of parameters. On the other hand, BIC *always* chose the correct models. Although in this model the components were well separated by construction, we found these results to be strong support for using these model selection criteria (especially BIC) in applications.

C.2   MONTE CARLO STUDY 2

We now describe a Monte Carlo study that investigated the effects of mixing distribution misspecification on model selection and parameter estimation. We use binomial denoninator 50 for all 27 observations in each replicate and the covariates $x_i = 0$ for $i \leq 18$ and $x_i = 1$ for $19 \leq i \leq 27$. Note that the covariates represent two groups, one having 18 observations and the other 9. Parameter values were chosen to be representative of the fitted models in Table 5. Although the model proposed in this article allows *both* regression coefficients and intercepts to vary across components, in order to allow comparison between slope estimates based on fitting a beta-binomial and a 2-component logistic mixed model, we base this study on models in which regression coefficients are constant but the intercept may vary between components.

The first model used in the study was a 2-component mixture in which the component mixing probabilities were .3 and .7 and the success probability functions had a common regression coefficient, $\alpha_1 = -.6$, and intercept values of $-.7$ and .2, respectively. Thus, the two success probability functions were logit$(-.7 - .6x_i)$ and logit$(.2 - .6x_i)$. The second model was a beta-binomial regression model in which the regression coefficient

again has a value of $\alpha_1 = -.6$, the intercept value was $-.05$ and the parameter $\phi = .2$ in the beta-binomial variance function (3.8).

For each model 500 replicates of 27 observations were generated. Data from each replicate was fit with both a beta-binomial model and a 2-component mixture model with constant regression coefficient. Summary statistics for the regression coefficient α_1 are reported in Table C.2. Observe that when the data were fit with the generating model, the average of the regression coefficient estimates were closer to the true value with smaller standard deviation. This limited analysis suggests that mixing distribution misspecification inflates both the bias and variability of regression coefficient estimates. Note also that standard deviations of parameter estimates were the largest when the mixed logistic model was fit to the beta-binomial data.

We also investigated the effect of using model selection criteria when the mixing distribution was misspecified. We used the AIC and BIC criteria to choose among one-to four-component discrete mixtures, with constant and variable regression coefficients for the data generated under the beta-binomial specification. Among the 500 cases the AIC criterion selected a 2-component model in 16 cases, a 3-component model in 262 cases, and a 4-component model in 222 cases, while the BIC selected a 2-component model in 27 cases, a 3-component model in 304, and a 4-component model in 169 cases. Hence, this simulation suggests that a mixture distribution with three or more support points would be required. Thus when a model with three or more components is suggested by the model selection criteria, a continuous mixing distribution might be a reasonable alternative to consider. In the context of the example in the previous section, this observation supports our use of a 2-component discrete mixture instead of a BBR model to describe the data.

ACKNOWLEDGMENTS

[Received October 1995. Revised February 1998.]

REFERENCES

Aitkin, M. A. (1995), "NPML Estimation in the Mixing Distribution in General Statistical Models With Unobserved Random Variation," in *Statistical Modelling*, eds. G. U. H. Seeber, B. J. Francis, R. Hatzinger, and G. Steckel-Berger, New York: Springer-Verlag.

Aitkan, M. A., and B. J. Francis (1995)," Fitting Overdispersed Generalized Linear Models by Nonparametric Maximum Likelihood," *GLIM Newsletter*, 37–45.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," Second International Symposium on Information Theory, eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kaido, 267–281.

Akaike, H. (1974), "A New Look At the Statistical Model Identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Anderson, D. A. (1988), "Some Models For Overdispersed Binomial Data", *Australian Journal of Statistics*, 30, 125–148.

Anderson, D. A., and M. Aitkin (1985), "Variance Component Models with Binary Response: Interviewer Variability," *Journal of the Royal Statistical Society,* Ser. B, 47, 203–210.

Blischke, W. R. (1964), "Estimating the Parameters of Mixtures of Binomial Distributions," *Journal of the American Statistical Association*, 59, 510–528.

Breslow, N. E., and Clayton, D. G. "Approximate Inference in Generalized Linear Mixed Models" *Journal of the American Statistical Association*, 88, 9–25.

Busvine, J. R. (1938), "The Toxicity of Ethylene Oxide to Calandra Oryzae, C. C. Granaria, Tribolium Castaneum, and Cimex Lectualarius," *Biology*, 25, 605–632.

Chatfield, C., and G. J. Goodhardt (1970), "The Betabinomial Model for Consumer Purchasing Behavior," *Applied Statistics*, 19, 240–250.

Chen, J., and J. D. Kalbfleisch (1996), " Penalized Minimum-Distance Estimates in Finite Mixture Models," *Canadian Journal of Statistics*, 24, 167–175.

Cox, D. R. (1983), "Some Remarks On Overdispersion," *Biometrika*, 70, 269–274.

Crowder, M. J. (1978), "Beta-Binomial ANOVA for Proportions," *Applied Statistics*, 27, 34–37.

Dean, C. (1992), "Testing for Overdispersion in Poisson and Binomial Regression Models," *Journal of the American Statistical Association*, 87, 451–457.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977),"Maximum Likelihood from Incomplete Data Via the EM Algorithm (with discussion)," *Journal of the Royal Statistical Society,* Ser. B, 39, 1–38.

Efron, B. (1986), "Double Exponential Families and Their Use in Gengeralized Linear Regression," *Journal of the American Statistical Association*, 81, 709–721.

Follmann, D. A., and Lambert, D. (1989), "Generalizing Logistic Regression by Nonparametric Mixing," *Journal of the American Statistical Association*, 84, 294–330.

Follmann, D. A., and Lambert, D. (1991), "Identifiability of Finite Mixture of Logistic Regression Models," *Journal of Statistical Planning and Inference,* 27, 375–381.

Griffiths, D. A. (1973), "Maximum Likelihood Estimation of the Betabinomial Distribution and an Application to the Household Incidence of the Total Number of Cases of a Disease," *Biometrics*, 29, 637–648.

Kahn, M. J., and Raftery, E. (1996), "Discharge Rates of Medicare Stroke Patients to Skilled Nursing Facilities: Bayesian Logistic Regression With Unobserved Heterogeneity," *Journal of the American Statistical Association*, 91, 29–41.

Laird, N. M. (1978), "Nonparametric Maximum Likelihood Estimation of Mixing Distribution," *Journal of the American Statistical Association*, 73, 805–811.

Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.

Lindsay, B. G. (1983), "The Geometry of Mixing Likelihood: A General Theory," *The Annals of Statistics*, 11, 86–94.

Margolin, B. H., Kim, B. S., and Risko, K. J. (1989), "The Ames Salmonella/Microsome Mutagenicity Assay: Issues of Inference and Validation," *Journal of the American Statistical Association* , 84, 651–661.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

McLachlan, G. J., and Basford, K. E. (1988), *Mixture Models*, New York: Marcel Dekker.

Nash, J. C. (1990), *Compact Numerical Methods for Computers* (2nd ed.), Bristol; New York: Adam Hilger.

Otake, M., and Prentice, R. L. (1984), "The Analysis of Chromosomally Aberrant Cells Based on Beta-Binomial Distribution," *Radiation Research*, 98, 456–470.

Pierce, D. A., and Sands, B. R. (1975), "Extra-Bernoulli Variation in Binary Data," Technical Report 46, Department of Statistics, Oregon State University, OR.

Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM*, 26, 195–239.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Teicher, H. (1961), "Identifiability of Mixtures," *The Annals of Mathematical Statistics*, 32, 244–248.

Titterington, D. M., Smith, A. F., and Markov, U. E. (1985), *Statistical Analysis of Finite Mixture Models*, Chichester: Wiley.

Wang, P. (1994), "Mixed Regression Models For Discrete Data," unpublished Ph.D. dissertation, University of British Columbia.

Wang, P., Puterman, M. L., Nhu, L., and Cockburn, I. (1996), "Mixed Poisson Regression Models With Covariate Dependent Rates," *Biometrics*, 52, 381–400.

Wang, P., Cockburn, I., and Puterman, M. L. (1998), "Patent Data Analysis—A Mixed Poisson Regression Model Approach," *Journal of Business and Economics Statistics*, 16, 27–41.

Williams, D. A. (1982), "Extra-binomial Variation in Logistic Linear Models," *Applied Statistics*, 31, 144–148.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.