MIXED REGRESSION MODELS FOR DISCRETE DATA

By

Peiming Wang

B. Sc. (Mathematics) Shanghai Second Polytechnic University ,1983

M. Sc. (Engineering) Shanghai Institute of Mechanical Engineering , 1988

M. A. (Statistics) York University, 1990

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
FACULTY OF
COMMERCE AND BUSINESS ADMINISTRATION

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

August, 1994

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Commerce

The University of British Columbia
Vancouver, Canada

Date Oct. 13, 1994

Abstract

The dissertation consists of two parts. In the first part we introduce and investigate a class of mixed Poisson regression models that include covariates in both mixing probabilities and Poisson rates. The proposed models generalize the usual Poisson regression in several ways, and can be used to adjust for extra-Poisson variation. The features of the models, identifiability, estimation methods based on the EM and quasi-Newton algorithms, properties of these estimates, model selection criteria and residual analysis are discussed. A Monte Carlo study investigates implementation and model choice issues. Several applications of this approach are analyzed. This analysis is compared to quasi-likelihood approaches.

In the second part we introduce and investigate a class of mixed logistic regression models that include covariates in both mixing probabilities and binomial parameters with the logit link. The proposed models generalize the usual logistic regression in several ways, and can be used to adjust for extra-binomial variation. The features of the models, identifiability, estimation methods based on the EM and quasi-Newton algorithms, properties of these estimates, model selection criteria and residual analysis are discussed. A Monte Carlo study investigates implementation and model choice issues. An applications of this approach is analyzed and results compared to those by quasi-likelihood approaches.

The dissertation also discusses future research in the areas and provides FORTRAN codes for all computations required to apply the models.

Table of Contents

List of Tables

List of Figures

Acknowledgement

First and foremost, I would like to thank my supervisor Prof. Martin L. Puterman. Marty suggested the basic models used in this thesis and provided me with much needed encouragement – especially in the early stages of our work. His assistance in the thesis research, writing and financial support through NSERC grant A5527 is deeply appreciated. His comments on drafts of thesis reflect a thoughtful serious reading and have substantially improved the final version. To him I give my deepest thanks.

Also, I would like to express my appreciation to my thesis committee members. Prof. Bent Jorgensen has raised many important questions about the thesis and has provided me with very valuable information on related research work. Prof. Iain Cockburn introduced me to patent data and provided helpful comments concerning econometrical issues related to it.

I also thank to Dr. Nhu Le for constructive comments concerning mixed Poisson regression models, and for providing assistance with seizure data analysis.

In addition, I acknowledge the receipt of MacPhee Memorial Fellowship and Leslie G. J. Wong Memorial Fellowship which provided support during my graduate schooling at this university.

Finally, I must thank my fellow students, faculty and staff of the management science division. The atmosphere has been open, relaxed and hospitable. I consider myself very fortunate to have known and become friends with so many people here.

To my parents

Chapter 1

Introduction

Poisson and logistic regression models are widely used for analyzing discrete data. Using such models, we implicitly assume that the response variable follows either a Poisson distribution or a binomial distribution with mean depending on covariates. Sometimes such assumptions may not be appropriate in the sense that the mean-variance relationship specified by the distribution of the response variable is not valid. In most of these cases, we often observe that data are overdispersed, i.e., the observed sample variance is larger than that predicted by inserting the observed sample mean into the mean-variance relationship. On the other hand, in few cases of data analysis, we may also observe that data are underdispersed, i.e., the observed sample variance is smaller than that predicted by inserting the observed sample mean into the mean-variance relationship. Without taking either overdispersion or underdispersion into account, using these regression models may lead to biased parameter estimates and incorrect inferences about the parameters. In this thesis, we propose using a finite mixture model approach to adjust for overdispersion. Specifically, we incorporate covariates in both mixing probabilities and component parameters of a finite mixture model in such a way that overdispersion may be explicitly interpreted by the model structure. The proposed models have applications in many different disciplines including economics, biostatistics and epidemiology.

The work in this thesis was motivated by several studies in different areas. One of these studies is to analyze relationship between technological innovation and research and development expenditures for U.S. high-tech companies. Another study is to assess

treatment effects in a clinical trial on epileptic patients carried out in British Columbia Children's Hospital. For the clinical study, for instance, the patients were randomly assigned into two groups: control and treatment. Those patients in the treatment group received monthly infusions of intravenous gammaglobulin (IVIG), while those patients in the control group received "best available therapy". The primary end point of the trial was daily seizure frequency. The principal data source was a daily seizure diary which contained the number of hours of parental observation and the number of seizures of each type during the observation period. We analyzed a typical series of myoclonic seizure counts from a single subject receiving IVIG. Data extracted from the seizure diary were the daily counts and the hours of parental observation. The questions of interest here are that of fitting a model to these counts which describes the pattern of epileptic seizure activity, and assessing IVIG effects on suppression of myoclonic seizures. Although it is a reasonable assumption that a daily seizure count follows a Poisson distribution which implies random occurrence of seizures in time, the data were overdispersed with respect to the Poisson regression model with mean including treatment effect. As indicated by the clinical investigators conducting this study, they have observed subjects to have "bad days" and "good days" with no obvious explanation of this effect. Hence, we are led to consider the mixed Poisson regression models which allow seizure frequency function to change in a random fashion.

Several alternative approaches for modelling overdispersion with respect to Poisson assumption are reviewed in Chapter 2. In this chapter, we propose a mixed Poisson regression model and show that it includes several special cases such as the usual Poisson regression model, mixed Poisson regression model with constant mixing probabilities and mixed Poisson regression model with constant Poisson rates. We also discuss identifiablity of the proposed model and provide sufficient conditions for identifiability. Maximum likelihood parameter estimation is used. An algorithm for computation of maximum

likelihood estimates is presented (FORTRAN code for implementation of the algorithm is provided in Appendix A). Particularly, for a fixed finite number of components, the algorithm finds maximum likelihood estimates by two steps: (1) using the EM algorithm first until either observed log likelihood or parameter estimates do not change more than a given tolerance, and (2) using a quasi-Newton algorithm which maximizes the observed log likelihood function. The results of a Monte Carlo study on performance of the algorithm are given here. Model selection procedure determining the number of components and inference about regression parameters is also presented. Classification based on the estimated posterior probabilities from the fitted model is discussed. Finally, four applications of this model are given, and results are compared to those from quasi-likelihood approaches.

Several alternative approaches for modelling overdispersion with respect to binomial assumption are reviewed in Chapter 3. In this chapter, we propose a mixed logistic regression model and show that it includes several special cases such as the usual logistic regression model, mixed logistic regression model with constant mixing probabilities and mixed logistic regression model with constant binomial parameters. We also discuss identifiablity of the proposed model and provide sufficient conditions for identifiability. Maximum likelihood parameter estimation is used. An algorithm for computation of maximum likelihood estimates is presented (FORTRAN code for implementation of the algorithm is provided in Appendix A). Particularly, for a fixed finite number of components, the algorithm finds maximum likelihood estimates by two steps: (1) using the EM algorithm first until either observed log likelihood or parameter estimates do not change more than a give tolerance, and (2) using a quasi-Newton algorithm maximizes the observed log likelihood function. The results of a Monte Carlo study on performance of the algorithm are given here. Model selection procedure determining the number of components and inference about regression parameters is also presented. Classification based

on the estimated posterior probabilities from the fitted model is discussed. Finally, an application of this model is given, and results are compared to those from quasi-likelihood approaches.

Chapter 4 concerns summary, conclusions and future research. We discuss some similarities and differences between the mixed Poisson regression and mixed logistic regression models. We extend the mixed Poisson regression and logistic regression models to the more general case of a one-parameter exponential distribution. Mixed exponential regression models are considered in this chapter. Furthermore, we propose hidden Markov Poisson regression models for longitudinal data. Particularly, we give preliminary results of this model, including model definition, moment structure, identifiability and parameter estimation.

Chapter 2

Mixed Poisson Regression Models

2.1 Poisson regression and its modifications

The Poisson regression model has been widely used for analyzing count data in which each observation consists of a discrete response variable and a vector of covariates or predictors. Typical examples of such data include counts of events in a Poisson or Poisson-like process where the upper limit to the number is infinite or effectively so. For instance, the response variable may represent the number of failures of a piece of equipment per unit time, the number of purchases of a particular commodity per family, or the number of bacteria per unit volume of suspension. In practice, however, the model sometimes fits poorly, suggesting the need for alternative models. In this case, it is not uncommon that observed data are overdispersed, i.e., the variance of an observation is greater than its mean. This may be reflected in over-large residual deviance and adjusted residuals which have a variance > 1. Without consideration for the overdispersion, using the Poisson regression model may not be justified. In the first part of this dissertation, mixed Poisson regression models are introduced and investigated. These models are applicable in several different situations where the Poisson regression model appears inadequate and provide an alternative way to adjust for extra-Poisson variation with a more meaningful interpretation.

Suppose that the ith response variable Y_i is a count, and associated with this response is a covariate vector $x_i = (x_{i1}, \ldots, x_{ir})'$ for $1 \leq i \leq n$. The Poisson regression model

assumes that the Y_i are distributed independently Poisson (λ_i) with density function

$$f(y_i \mid \alpha, x_i) = \frac{1}{y_i!}\lambda_i^{y_i} \exp(-\lambda_i) \quad \text{for } y_i = 0, 1, 2, \ldots, \tag{2.1}$$

where $\lambda_i = \exp(x_i'\alpha)$, $\alpha \in R^r$ is a r-dimensional vector of unknown parameters. Note that the Poisson parameter $\lambda_i = E(Y_i)$ is related to the covariate vector x_i by a link function so that the dependence of λ_i on x_i is assumed to be multiplicative and is usually written in the logarithmic form

$$\log(\lambda_i) = x_i'\alpha. \tag{2.2}$$

Equations (2.1) and (2.2) are sometimes referred as a log-linear model.

The Poisson regression model has been applied in many areas (e.g., Frome, Kutner, and Beauchamp 1973; Frome 1983; Holford 1983; Hausman et al. 1984; Mannering 1989). For instance, Frome et al (1973) used the Poisson regression model to describe the relationship between the number of failures of a piece of electronic equipment per unit time (response variable) and the times spent in regimes one and two (covariates), and the relationship between the number of colonies produced in the spleen of recipient animals (response variable) and the concentration of injected cells and the radiation dose (covariates). Frome (1983) applied the Poisson regression model in the analysis of survival time data. He analyzed the data that were obtained in epidemiologic follow-up studies and organized into a format similar to that of a life table. Holford (1983) analyzed the data that consists of numbers of prostatic cancer deaths and mid-period population denominators for non-whites in the US by age and calendar period, and fitted it to the Poisson regression model with age and cohort effects to the death rates. Hausman et al. (1984) introduced the Poisson regression model to analyze the relationship between the research and development (R&D) expenditures of firms and the number of patents applied for and received by them. Mannering (1989) used the Poisson regression model to investigate

the determinants of commuter flexibility in changing routes and departure times for the morning trip to work. He assumed that the number of route and departure time changes occurring during a one month period follows a Poisson distribution with mean depending on a vector of commuting and socioeconomic characteristics for an individual.

The Poisson regression model is analogous to the normal linear regression model in many ways. The estimation of unknown parameters is straightforward and is done either by an iterative weighted least squares technique or by a maximum likelihood algorithm. The log likelihood function is globally concave so that maximization routines converge rapidly. Residual analysis is carried out in the same way as the normal linear regression model, except that the definition of the residual is different.

The Poisson regression model is used for many different purposes. Sometimes, inference concerning the regression parameters α is of primary importance. For example, Y may denote the number of car accidents for an individual. Large values of αs (relative to their standard errors) then correspond to factors which significantly increase the chance of the accidents. On the other hand, when one is primarily interested in creating a good predictive model, the interpretation of parameters may take a secondary role.

The Poisson regression model is an example of a Generalized Linear Model (McCullagh and Nelder, 1989) in which the frequency distribution of the response Y is a Poisson distribution with mean $\lambda(x)$, and the link is a log function: $g(\lambda) = \log(\lambda(x)) = x'\alpha$.

A consequence of using the Poisson regression model is that the variance equals the mean, i.e., $Var(Y_i) = E(Y_i)$. In practice, however, we often have overdispersed data, i.e., $Var(Y_i) > E(Y_i)$. When the Poisson regression model fits the count data poorly, overdispersion is often a cause of the problem. There are several ways to modify the Poisson regression model. Using GLM formulation we can modify it by choosing either an alternative link function or an alternative frequency distribution, or both. Since the log link has nice properties such as multiplicative effects of covariates on the Poisson

mean, few researchers have suggested use of alternative link functions. On the other hand, there are a lot of studies of alternative frequency distributions for the Poisson distribution (e.g., Breslow 1984; Efron 1986; Lawless 1987b and Dean et al. 1989).

To adjust for extra-Poisson variation, mixed Poisson distributions have been used as frequency distributions (Efron 1986; Lawless 1987b and Dean et al. 1989). In these models, the Poisson means associated with each observed count are defined as latent variables that are sampled from a specified parametric distribution. In other words, the Poisson means are random variables following a specific distribution. Under such a set-up, the marginal density function of the response Y without covariates can be often given by

$$Pr(Y = y \mid \lambda, g) = \int_0^\infty \frac{1}{y!}[v\lambda]^y \exp(-v\lambda) g(v) dv, \quad \text{y=0,1,...} \tag{2.3}$$

where $g(v)$ is a mixing probability density function and $\lambda > 0$ is a unknown parameter. Such models can be viewed as multiplicative Poisson random-effects models (Brillinger 1986) for the following reasons: (1) there is a random effect Υ with a density $g(v)$, $v > 0$ in the model; (2) conditional on $\Upsilon = v$, the response Y has a Poisson distribution with mean $v\lambda$. Without loss of generality we can assume that $E(\Upsilon) = 1$.

Most authors have considered a gamma mixing distribution, which leads to a negative binomial distribution for the observed data (Manton, Woodbury, and Stallard 1981, Margolin, Kaplan, and Zeiger 1981). In this case the mixing distribution $g(v)$ is

$$g(v) = \begin{cases} \frac{k^k}{\Gamma(k)} v^{k-1} \exp(-kv) & \text{for } v \geq 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

where $k > 0$ and $\lambda > 0$ are unknown parameters. Note that $E(\Upsilon) = 1$ and $Var(\Upsilon) = 1/k$. Hence (2.3) becomes

$$f(y \mid \lambda, k) = \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{\lambda}{k+\lambda}\right)^y \left(\frac{k}{k+\lambda}\right)^k, \quad \text{for } y = 0, 1, 2, \ldots, \tag{2.4}$$

where $k \geq 0$ is often referred to the index or dispersion parameter. The mean and variance of Y are

$$E(Y) = \lambda \quad \text{and} \quad Var(Y) = \lambda + (1/k)\lambda^2. \tag{2.5}$$

As a natural extension of the above models, several researchers (e.g., Lawless, 1987b, and Hausman, Hall, and Griliches, 1984) have studied negative binomial regression models in which covariates are related to the parameter λ by a positive function $\lambda(x)$. Usually one takes the common log-linear form $\lambda(x) = \exp(x'\alpha)$ so that random and fixed effects are added on the same exponential scale. The negative binomial regression model may be interpreted as follows: if Υ is a positive-value random variable with mean 1 and variance $1/k$, and if the distribution of Y, conditional on $\Upsilon = v$ and covariates x, is Poisson $(v\lambda(x))$, then the marginal mean and variance of Y are as in (2.5), and the marginal distribution of Y is the negative binomial defined by (2.4).

Note that in the negative binomial regression model, the shape parameter k is a constant for all observations. In this case, the likelihood equations based on the negative binomial model are unbiased and the maximum likelihood estimates of the mean parameters are consistent, regardless of the true variance function (Lawless, 1987b and Hausman, Hall, and Griliches 1984).

Several researchers apply the negative binomial model in different situations. For instance, for count data without covariates, Anscombe (1950) gives a comprehensive discussion of properties of the model and several examples of the use of the model. Ehrenberg (1972) applies it to model market behaviour for frequently purchased low-cost products by assuming that the number of purchases follows the negative binomial distribution. For count data with covariates, Manton et al. (1981) use it in the analysis of mortality rates. They assume that variation in individual risk levels follows the gamma distribution within each category, and that conditional on the individual risk levels, the

number of cancer deaths follows the Poisson distribution with mean depending on some covariates including age and race. Hausman, Hall, and Griliches (1984) introduce it to study the relation between technical innovation and firm characteristics (mainly R&D spending and sales) at firm level. They assume that there is a random firm effect described by the gamma distribution, and that number of patents applied for by a company per year, Y, follows a negative binomial regression model in which $E(Y) = \lambda(x)$ is a log-linear function of the covariates: annual R&D spending and sales of the company.

Another useful choice of the mixing distribution $g(v)$ in (2.3) is an inverse Gaussian distribution (e.g., Folks and Chhikara 1978, Tweedie 1957) for Υ, with density

$$g(v) = (2\pi\tau v^3)^{-1/2} \exp(-(v-1)^2/2\tau v), \quad v > 0. \tag{2.6}$$

The parameter τ is unknown, and equals $Var(\Upsilon)$. The marginal distribution of Y from (2.3) is then a Poisson-inverse-Gaussian model with the mean and variance relationship: $E(Y) = \lambda$ and $Var(Y) = \lambda + \tau\lambda^2$. This model provides a heavier-tailed alternative to the negative-binomial model, although both have the same mean and variance relationship. A difficulty of using the model is to compute the integral in (2.3).

Dean, Lawless and Willmot (1989) introduce a Poisson-inverse-Gaussian regression model by taking the common log-linear form $\lambda(x) = \exp(x'\alpha)$. This model has almost the same structure and interpretation as the negative binomial regression models. Jorgensen (1987) and Stein and Juritz (1988) also propose other versions of Poisson-inverse-Gaussian models by using different variance functions. Jorgensen (1987) defines both the Poisson and inverse-Gaussian distributions as exponential dispersion models so that his mixture model is an exponential dispersion model and satisfies an appealing convolutions property. Stein and Juritz's model is structured so that the regression parameter vector α is orthogonal to the shape parameter (analogous to the τ in the above model) specifying the degree of extra-Poisson variation. Neither model has, however, the simple structure

of the above model in terms of the multiplicative random effects.

A log normal mixing distribution for $g(v)$ has also been advocated (e.g., Hinde 1982 and Pocock et al 1981). In this model, the Poisson mean has a lognormal distribution with location parameter related to a linear function of covariates and a constant scale parameter.

Efron (1986) introduces the double Poisson distribution as an alternative frequency distribution to accommodate extra-Poisson variation. The exact double Poisson density is

$$\tilde{f}_{\lambda,\theta}(y) = c(\lambda,\theta) f_{\lambda,\theta}(y),$$

where

$$f_{\lambda,\theta}(y) = (\theta^{1/2} e^{-\theta\lambda}) \left(\frac{e^{-y} y^y}{y!}\right) \left(\frac{e\lambda}{y}\right)^{\theta y}, \quad \text{for } y = 0, 1, 2, \ldots,$$

and the factor $c(\lambda,\theta)$ can be calculated as

$$\frac{1}{c(\lambda,\theta)} = \sum_{y=0}^{\infty} f_{\lambda,\theta}(y) \approx 1 + \frac{1-\theta}{12\lambda k}(1 + \frac{1}{\lambda\theta}).$$

Since the constant $c(\lambda,\theta)$ nearly equals 1, the approximate probability density function for the double Poisson distribution is $f_{\lambda,\theta}(y)$. Usually λ is referred to as a mean parameter and θ as a dispersion parameter. The double Poisson distribution allows us to individually adjust the mean and variance of the response Y using the parameters λ and θ, and it only involves rescaled Poisson distributions, in the approximate sense that Y is approximately expressed by X/θ where X follows the Poisson distribution with mean $\lambda\theta$. For count data with covariates, we can incorporate covariates to either λ or θ or both. Efron suggests that the double Poisson regression model may be more appropriate for count data in which subjects may be, for example, obtained in clumps rather than by genuine random sampling. Note that such clumped sampling may be one of possible causes of overdispersion.

Another approach to modify the Poisson regression distribution is the quasi-likelihood. This approach specifies only the mean and variance structure of Y implied by the mixed Poisson model, and estimates the regression coefficients by quasi-likelihood and the variance parameter by the method of moments (e.g., Williams 1982 and Breslow 1987). The attraction is that unduly rigorous assumptions about the frequency distribution are avoided. The trade-off is that the estimation based on the quasi-likelihood model is not as efficient as the fully parametric model (Lawless 1987b).

Several researchers have studied different quasi-likelihood models by assuming different relationship between mean and variance. Breslow (1984) introduces the quasi-likelihood models by assuming that conditional on λ_i and exposure t_i, the response Y_i has an independent Poisson distribution with mean $E(Y_i) = \lambda_i t_i$, and $\log(\lambda_i) = x_i'\alpha + \epsilon_i$ where α is a vector of unknown parameters and the ϵ_i are random error terms having means 0 and a constant unknown variance σ^2. Note that there are no assumptions on the probability distributions of random effects ϵ_i except the first two moments.

Breslow (1984) also proposes two procedures to fit count data to the model. One is when the data have relatively large values of Y_i. In this case $Z_i = \log(Y_i/t_i)$ may be regarded as having approximate normal distributions with mean $x_i'\alpha$ and variance $\sigma^2 + \tau_i^2$ where $\tau_i^2 = 1/E(Y_i)$. Hence the estimation method is based on the iteration of the following two steps: (1) obtain estimates of the regression parameters by weighted least squares solution using the empirical weights $w_i = (\sigma^2 + \hat{\tau}_i)^{-1}$, and (2) obtain the value of σ^2 by setting the chi-square criterion equal to its degree of freedom, i.e.,

$$\sum_{i=1}^{n}(z_i - x_i'\alpha)^2/(\sigma^2 + \tau_i^2) = n - p,$$

where p is the number of parameters in the model.

The other is when the data have relatively small values of Y_i. In this case, the normal approximation appears in doubt. Since the above assumptions lead to the approximate

mean and variance relationship: $E(Y_i) = t_i\lambda_i \simeq \exp(x_i'\alpha)$ and $Var(Y_i) \simeq \lambda_i + \sigma^2\lambda_i^2$, the maximum quasi-likelihood estimates are obtained with GLIM (Backer and Nelder, 1978) by using Poisson error function and the natural log link, declaring $\log(t_i)$ as an offset, and defining prior weights $w_i = (1 + \sigma^2\hat{\lambda}_i)^{-1}$. The value of σ^2 is also obtained by setting the chi-square criterion equal to its degrees of freedom, i.e.,

$$\sum_{i=1}^{n}(y_i - \hat{\lambda}_i)^2/\{\lambda_i(1 + \sigma^2\hat{\lambda}_i)\} = n - p$$

where p is the number of parameters in the model. Note that this approach can also apply to the data that have both small and large values of Y_i because the above approximation of the mean and variance relationship can still hold.

There are also other quasi-likelihood models in the literature for analyzing overdispersed count data. For instance, many non-Poisson distributions encountered in statistical practice may have the connection between the mean and variance of a response Y as expressed by

$$Var(Y) = c_1E(Y) + c_2\{E(Y)\}^2.$$

This relation was used by Bartlett (1936) to analyze counts for field experiments. Both Armitage (1957) and Finney (1976) define another mean-variance relationship as

$$Var(Y) = c\{E(Y)\}^b,$$

and find by the study of examples that $1 < b < 2$. Breslow (1990) also uses a quasi-likelihood model with the above mean-variance relationship to analyze viral activity from pock counts.

Another approach for modifying the Poisson distribution is through finite mixture

models which are obtained by taking the mixing distribution in (2.3) as a discrete probability distribution with c points of support. Hence the distribution of Y is

$$Pr(Y = y \mid p_1, \ldots, p_c, \lambda_1, \ldots, \lambda_c) = \sum_{j=1}^{c} p_j \mathrm{Po}(y \mid \lambda_j),$$

where $\sum_{j=1}^{c} p_j = 1$ and $p_j > 0$ $(1 \leq j \leq c)$, and $\mathrm{Po}(y \mid \lambda_j)$ are Poisson distribution functions with mean λ_j. This approach applies to a wide variety of applications and has received an increasing amount of attention late. See for example Everitt and Hand (1981) and Titterington et al (1985). Simar (1976) and Leroux (1989) study finite mixtures with an unknown number components for overdispersed count data. No researchers have systematically studied regression-type finite mixture models with covariates.

2.2 Implications of Overdispersion

Overdispersion as an issue has been recognized for many years. In Poisson regression analysis of count data, residual variability sometimes is greater than what is predicted by Poisson models, suggesting either lack-of-fit (incorrect mean) or overdispersion, or both. It is important to note that so far various score tests cannot distinguish lack-of-fit from the true overdispersion (incorrect variance). In our discussion, we mainly concentrate on the issue of overdispersion rather than the choice of the link function. Without consideration of overdispersion, using the Poisson regression model may be misleading in statistical analysis. This will be illustrated in our examples later.

Many authors have studied the effects of overdispersion on inferences made under the Poisson regression model. As Cox (1983) indicates, overdispersion in general has two effects. One is that summary statistics have a larger variance than anticipated under the simple model. The second effect is a possible loss of efficiency. It is important to note that the implications of overdispersion may also depend on the type of overdispersion specified. For the Poisson regression analysis, if the overdispersion is accommodated by randomizing

the Poisson mean to obtain gamma-Poisson models and quasi-likelihood models, among others (e.g. Cox 1983), fitting maximum likelihood of a log linear model for Poisson-distribution data retains high efficiency for a modest amount of overdispersion, provided that the log linear model determines the expected value of the observed count (Cox, 1983). Specifically parameter estimates based on the Poisson regression model are generally not seriously biased or inefficient, but estimated standard errors are too small and tests are too liberal (Breslow 1990; Cox 1983; Firth 1987; Hill and Tsai 1988; McCullagh and Nelder 1989).

On the other hand, when there is serious overdispersion, using the usual Poisson regression may lead to either seriously biased or inefficient parameter estimates. For instance, in a random coefficient log-linear Poisson regression, the response Y is Poisson $(e^{\alpha+x\beta})$ given α and β, but each individual has a different random baseline α or different responsiveness to treatment β, parameter estimates of α and β as well as their standard errors based on the Poisson regression may be misleading. In particular, the mean of a random coefficient is not the Poisson mean evaluated at the average of the random coefficients (see Neuhaus et al. 1991). Also if the true log-mean is $\alpha + x\beta + z\gamma$ but only x is recorded, then the assumed log-mean $\alpha^* + x\beta$ has a random intercept $\alpha^* + z\gamma$ that varies with z. In this case the extent of the overdispersion depends on z, and the parameter estimate of α^* based on the Poisson regression may be seriously biased when the overdispersion is serious. When the extra-Poisson variation is explained by the mixed Poisson regression model, we will show, in examples, that without accounting for the overdispersion may have rather different results from the usual Poisson regression.

2.3 Tests for Extra-Poisson Variation

There are several overdispersed Poisson regression models which have been discussed in the literature. Without fitting a particular overdispersed Poisson model, we would like to know whether there is serious overdispersion. Several methods have been proposed to detect overdispersion in terms of the Poisson assumption. An informal graphical approach is introduced by Lambert and Roeder (1993) and Lindsay and Roeder (1992). For instance, for log-linear Poisson regression, Lambert and Roeder (1993) define the following function

$$C(\mu) = n^{-1} \sum_{i=1}^{n} \exp(\hat{y}_i - \mu) \left(\frac{\mu}{y_i}\right)^{y_i},$$

where $\hat{y}_i = \exp(x_i'\hat{\beta})$ and $\mu > 0$. They show that $C(\mu)$ tends to be convex when the data are from a random mean Poisson regression model, random coefficient Poisson regression model, or double Poisson regression model. Thus they suggest to use the plot of $C(\mu)$ against μ. The more convex $C(\mu)$ appears, the more evidence there is of overdispersion or an omitted variable. It is not clear, however, whether this approach can apply to other modified Poisson regression models such as the finite mixture of Poisson regression model for dealing with extra-Poisson variation.

Another simple approach is to fit a more comprehensive model that contains the Poisson model and then test for a reduction to the simple model using, for instance, a likelihood ratio test. This approach, however, may provide misleading results (Dean, 1992). As Lawless (1987a) indicates, in certain circumstances the asymptotic distributions used with these tests may not be reliable because they tend to underestimate the evidence against the base model.

A widely used approach is through score tests. With these tests we may fit the Poisson

regression model as a first step in the model building process and test for overdispersion. Score tests for detecting extra-Poisson variation have been discussed by Cameron and Trivedi (1986), Collings and Margolin (1985), Dean and Lawless (1989), and Fisher (1950). Concern has been expressed over the suitability of tests and confidence interval based on overly simple models for extra-Poisson. Breslow (1990) proposes tests for parameters that appear in the mean, using model-free estimates of variance for each case. He found that these to be robust to incorrect specification of the variance function, but not as powerful as tests based on correct model for response variation. Dean (1992) develops a unifying theory for all the score tests mentioned above.

Before applying the mixed Poisson regression models, we need to determine whether the data are overdispersed with respect to the Poisson distribution in Poisson regression models. We use three score test statistics proposed by Dean (1992). They test the hypothesis of no overdispersion against alternatives representing different forms of overdispersion. The test statistics are

$$\begin{aligned} P_a &= \frac{\sum((y_i - \hat{\mu}_i)^2 - \hat{\mu}_i)}{\sqrt{2\sum \hat{\mu}_i^2}}, \\ P_b &= \frac{\sum((y_i - \hat{\mu}_i)^2 - y_i)}{\sqrt{2\sum \hat{\mu}_i^2}}, \\ \text{and} \quad P_c &= \frac{1}{\sqrt{2n}} \sum \frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} \end{aligned}$$

corresponding to the following specifications of overdispersion:

(a) $E(y_i) \simeq \mu_i$, $Var(y_i) \simeq \mu_i(1 + \tau\mu_i)$ for τ small;

(b) $E(y_i) = \mu_i$, $Var(y_i) = \mu_i(1 + \tau\mu_i)$;

(c) $E(y_i) = \mu_i$, $Var(y_i) = \mu_i(1 + \tau)$.

In these formulae $\hat{\mu}_i$ is the estimated mean value for the independent identical observations based on Poisson regression. Under H_0: $\tau = 0$, each asymptotically follows a standard normal distribution. Note that the difference between (a) and (b) is that the

former has the approximate forms for the first two moments, whereas the latter has the exact ones.

For small samples, Dean (1992) provides the following "corrected" versions P_a', P_b' and P_c' corresponding to P_a, P_b and P_c respectively.

$$\begin{aligned} P_a' &= \frac{\sum((y_i - \hat{\mu}_i)^2 - (1 - \hat{h}_{ii})\hat{\mu}_i)}{\sqrt{2\sum \hat{\mu}_i^2}}, \\ P_b' &= \frac{\sum((y_i - \hat{\mu}_i)^2 - y_i + \hat{h}_{ii}\hat{\mu}_i)}{\sqrt{2\sum \hat{\mu}_i^2}}, \\ \text{and} \quad P_c' &= \frac{1}{\sqrt{2n}} \sum \frac{(y_i - \hat{\mu}_i)^2 - y_i + \hat{h}_{ii}\hat{\mu}_i}{\hat{\mu}_i} \end{aligned}$$

where $\hat{h}_{ii}$ is the ith diagonal element of the matrix $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$, with $W = diag(\hat{\mu}_1, \ldots, \hat{\mu}_n)$ and X being an $n \times p$ design matrix. Dean (1992) points out that the distributions of these corrected statistics converges very quickly to normality.

2.4 Mixed Poisson Regression Models

Without covariates, the finite mixture approach has been used for analyzing count data appearing extra-Poisson variation (c.f. Titterington et al.,1985; Simar 1976; and Leroux, 1989). With covariates, however, this approach has not been systematically studied and directly applied for analyzing regression-type count data. In this section, we extend the finite Poisson mixture model to the mixed Poisson regression model by allowing both the component Poisson parameters and mixing probabilities of a mixture to depend on covariates. We investigate some basic features of the model. We also discuss identifiability for the model and provide sufficient conditions for the identifiability.

2.4.1 The Model

Let the random variable Y_i denote the ith count response, and let $\{(y_i, t_i, x_i),\ i = 1, \ldots, n\}$ denote observations where y_i is the observed value of Y_i, t_i a non-negative number representing the time period or exposure during which observation y_i is generated, and $x_i = (x_i^{(r)}, x_i^{(m)})$ a covariate vector in which $x_i^{(r)}$ and $x_i^{(m)}$ are k_1 and k_2-dimensional covariate vectors corresponding to the regression part and the mixing part of the model respectively. We allow some or all components of $x_i^{(r)}$ and $x_i^{(m)}$ to be identical. Usually the first elements of $x_i^{(m)}$ and $x_i^{(r)}$ is a 1 corresponding to an intercept. The mixed Poisson regression model assumes

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

(2) For each observed count y_i, there is an unobserved random variable, Λ_i, representing the component which generates y_i. Further, the (Y_i, Λ_i) are pairwisely independent;

(3) Λ_i follow discrete distributions with c points of support, $1, \ldots, c$, and

$$Pr(\Lambda_i = j) = p_{ij},$$

where $\sum_{j=1}^{c} p_{ij} = 1$ for each i and

$$\begin{aligned} p_{ij} &\equiv p_j(x_i^{(m)}, \beta) \\ &= \frac{\exp(\beta_j' x_i^{(m)})}{1 + \sum_{k=1}^{c-1} \exp(\beta_k' x_i^{(m)})} \quad \text{for } j = 1, \ldots, c-1, \text{ and} \end{aligned} \tag{2.7}$$

$$p_{ic} \equiv p_c(x_i^{(m)}, \beta) = 1 - \sum_{j=1}^{c-1} p_{ij}, \tag{2.8}$$

where $\beta = (\beta_1, \ldots, \beta_{c-1})'$ and $\beta_j = (\beta_{j1}, \ldots, \beta_{jk_2})'$, $j = 1, \ldots, c-1$, are unknown parameters. Note that *all* components of β appear in each mixing probability p_{ij}, $j = 1, \ldots, c$;

(4) Conditional on $\Lambda_i = j$, Y_i follows a Poisson distribution which we denote by

$$\begin{aligned} Y_i &\sim f_j(y_i \mid x_i^{(r)}, t_i, \alpha_j) \\ &\equiv \mathrm{Po}(y_i \mid \lambda_{ij}) \\ &= \frac{1}{y_i!}\lambda_{ij}^{y_i}\exp(-\lambda_{ij}) \end{aligned} \tag{2.9}$$

where we define a log link function between the Poisson mean and covariates as

$$\lambda_{ij} \equiv t_i\lambda_j(x_i^{(r)}, \alpha_j) \equiv t_i \exp(\alpha_j' x_i^{(r)}), \quad \text{for } j = 1, \ldots, c,$$

where $\alpha \equiv (\alpha_1, \ldots, \alpha_c)'$ are unknown parameters , and $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jk_1})'$, $j = 1, \ldots, c$. Note that we could also choose other link functions.

The above assumptions define the unconditional distribution of observations, y_i, as a finite Poisson mixture in which the mixing probabilities, p_{ij}, are related to the covariates $x_i^{(m)}$ through the logit function, and the component distributions are Poisson distributions with mean determined by the exposure, t_i and by the Poisson rate $\lambda_j(x_i^{(r)}, \alpha_j)$, which is related to the covariates $x_i^{(r)}$ through an exponential function. Suppose that observations can be classified into c groups corresponding to the c underlying states, a vector of unknown parameters α_j may be interpreted as the coefficients of the Poisson regression for group j. On the other hand, unknown parameters β may be interpreted as the coefficients of the multinomial regression in which Λ_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 of Poisson rates, αs, to be constant across components, i.e., $\alpha_{jl} = \alpha_l$ for $j = 1, \ldots, c$ or 0 in one or several covariates, i.e., $\alpha_{jl} = 0$ for some j, $j = 1, \ldots, c$.

Under the above assumptions the probability function of Y_i satisfies

$$f(y_i \mid x_i^{(r)}, x_i^{(m)}, t_i, \alpha, \beta) = \sum_{j=1}^{c} p_{ij}\mathrm{Po}\,(y_i \mid \lambda_{ij}) \tag{2.10}$$

where p_{ij} and $\mathrm{Po}\,(y_i \mid \lambda_{ij})$ are given by (2.7), (2.8) and (2.9) 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 probability p_{ij}; Subsequently, y_i is generated from a Poisson distribution with mean λ_{ij}.

A justification for the mixed Poisson regression models is to assume that the coefficient vector α in the usual Poisson regression model, $\log(\lambda_i) = \alpha' x_i^{(r)}$, 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 p_j are related to a covariate vector $x_i^{(m)}$ through a logit link $p_j(x_i^{(m)}, \beta)$ we are led to the model of equation (2.10).

Note that this model includes many previously studied models as special cases.

- Choosing $c = 1$ yields the Poisson regression model;
- Setting $x_i^{(r)} = x_i^{(m)} = 1$ and $t_i = 1$ for all i yields an independent Poisson mixture model (Simar (1976) and Leroux (1989));
- Setting $x_i^{(m)} = 1$ yields an independent finite mixture of Poisson regression. Further, letting the Poisson rates have common regression parameters and different intercepts yields a Poisson regression with a random intercept which follows a discrete mixing distribution;
- Setting $x_i^{(m)} = 1$, $c = 2$ and $\lambda_1(x_i^{(r)}, \alpha_1) \equiv 0$ yields a Poisson regression model with an extra mass at 0;
- Setting $x_i^{(r)} = 1$ yields an independent multinomial mixture of Poisson distributions with constant rates.

For the above model, the mean and variance of observation y_i are, respectively,

$$E(Y_i) \;=\; E(E(Y_i \mid \Lambda_i))$$

$$= t_i \sum_{j=1}^{c} p_{ij}\lambda_{ij} \tag{2.11}$$

and

$$\begin{aligned} Var(Y_i) &= E(Var(Y_i \mid \Lambda_i)) + Var(E(Y_i \mid \Lambda_i)) \\ &= t_i \sum_{j=1}^{c} p_{ij}\lambda_{ij} + t_i^2 \left\{ \sum_{j=1}^{c} p_{ij}\lambda_{ij}^2 - \left\{ \sum_{j=1}^{c} p_{ij}\lambda_{ij} \right\}^2 \right\} \end{aligned} \tag{2.12}$$

Obviously, $Var(E(Y_i \mid \Lambda_i)) = 0$ if and only if

$$\lambda_{i1} = \lambda_{i2} = \ldots = \lambda_{ic}. \tag{2.13}$$

This implies that the mixture model is able to cope with extra-Poisson variation among $Y_1, \ldots, Y_n$ due to heterogeneity in the population.

2.4.2 Identifiability

To be able to estimate the parameters of (2.10), it is important to establish identifiability of the model, that is, two sets of parameters in the mixture which do not agree after permutation cannot yield the same mixture distribution. Furthermore, identifiability is a necessary requirement for the usual asymptotic theory to hold for the estimation procedures considered latter. For finite mixture models with covariates we define identifiability as follows.

Let $\mathcal{F} = \{F(x, \theta); \theta \in \Sigma, x \in R^d\}$ be the class of d-dimensional distribution functions from which finite mixtures are to be formed. This class is *identifiable* if

$$\sum_{j=1}^{c} p_j F_j(x) = \sum_{j=1}^{\tilde{c}} \tilde{p}_j \tilde{F}_j(x) \quad \text{for } x \in R^d,$$

where $\sum_j^c p_j = \sum_j^{\tilde{c}} \tilde{p}_j = 1$ and $p_j, \tilde{p}_j$ are positive, implies that $c = \tilde{c}$ and we can order the summations such that $p_j = \tilde{p}_j$, $F_j = \tilde{F}_j$, $j = 1, \ldots, c$. Note that if a class of models

is not identifiable we cannot discriminate between (at least two) parameter values using data generated by the model.

Without covariates, Teicher (1961) proves that the class of finite mixtures of Poisson distributions is identifiable. Considering covariates, we extend the above definition of identifiability as follows.

Definition 1: Consider the collection of probability models $\{f(y_1 \mid x_1^{(r)}, x_1^{(m)}, t_1, \alpha, \beta), \ldots, f(y_n \mid x_n^{(r)}, x_n^{(m)}, t_n, \alpha, \beta)\}$, with a restriction that $\lambda_{1i} < \ldots < \lambda_{1c}$, parameter space $\mathcal{C} \times \Lambda \times \mathcal{P}$, sample spaces $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$, and fixed covariate vectors $(x_1^{(r)}, x_1^{(m)}), \ldots, (x_n^{(r)}, x_n^{(m)})$ where $x_i^{(r)} \in R^{k_1}$ and $x_i^{(m)} \in R^{k_2}$ for $i = 1, \ldots, n$. The collection of probability models is *identifiable* if for $(c, \alpha, \beta), (c^*, \alpha^*, \beta^*) \in \mathcal{C} \times \Lambda \times \mathcal{P}$,

$$f(y_i \mid x_i^{(r)}, x_i^{(m)}, t_i, \alpha, \beta) = f(y_i \mid x_i^{(r)}, x_i^{(m)}, t_i, \alpha^*, \beta^*) \tag{2.14}$$

for all $y_i \in \mathcal{Y}_i$, $i = 1, \ldots, n$, implies $(c, \alpha, \beta) = (c^*, \alpha^*, \beta^*)$.

Note that the order restriction in the definition means that two models are equivalent if they agree up to permutations of parameters.

We now provide sufficient conditions for identifiability.

Theorem 1: The mixed Poisson regression model is identifiable if both matrices $X^{(m)}$ and $X^{(r)}$ are full rank, where $X^{(m)} = (x_1^{(m)} x_2^{(m)} \ldots x_n^{(m)})'$ and $X^{(r)} = (x_1^{(r)} x_2^{(r)} \ldots x_n^{(r)})'$.

Proof: Suppose that $(c, \alpha, \beta), (c^*, \alpha^*, \beta^*)$ satisfy (2.14). This then implies that for each i and all $y_i \in \mathcal{Y}_i$, $i = 1, \ldots, n$,

$$\sum_{j=1}^{c} p_{ij} \mathrm{Po}(y_i \mid \lambda_{ij}) = \sum_{j=1}^{c^*} p_{ij}^* \mathrm{Po}(y_i \mid \lambda_{ij}^*), \tag{2.15}$$

where $p_{ij} = p_j(x_i^{(m)}, \beta)$ and $\lambda_{ij} = \lambda_j(x_i^{(r)}, \alpha_j)$ are defined above. Note that each side of (10) may be regarded as a finite Poisson mixture without covariates. Teicher's result implies that

$$c = c^*, \quad p_{ij} = p_{ij}^* \quad \text{and} \quad \lambda_{ij} = \lambda_{ij}^*$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, c$. By the definition of the model, we obtain

$$\exp(\beta_j' x_i^{(m)}) = \exp(\beta_j^{*\prime} x_i^{(m)}) \text{ for } j = 1, \ldots, c-1 \tag{2.16}$$

$$\exp(\alpha_j' x_i^{(r)}) = \exp(\alpha_j^{*\prime} x_i^{(r)}) \text{ for } j = 1, \ldots, c \tag{2.17}$$

From (2.16) and (2.17) we obtain

$$(\beta_j - \beta_j^*)' x_i^{(m)} = 0 \text{ for } j = 1, \ldots, c-1 \text{ and } i = 1, \ldots, n$$

$$(\alpha_j - \alpha_j^*)' x_i^{(r)} = 0 \text{ for } j = 1, \ldots, c \text{ and } i = 1, \ldots, n$$

or

$$(\beta_j - \beta_j^*)' X^{(m)} = 0 \text{ for } j = 1, \ldots, c-1 \tag{2.18}$$

$$(\alpha_j - \alpha_j^*)' X^{(r)} = 0 \text{ for } j = 1, \ldots, c. \tag{2.19}$$

Sufficient conditions for identifiability are that both $X^{(m)}$ and $X^{(r)}$ are full rank matrices, in which case (2.18) and (2.19) imply that $(\alpha, \beta) = (\alpha^*, \beta^*)$. We can assume this without loss of generality such as might be the case in an ANOVA structure, since if it does not hold we can reparameterize the model accordingly. □

2.5 Parameter Estimation for the mixed Poisson regression models

To find the maximum likelihood estimates of the parameters in the mixed Poisson regression model requires an iterative algorithm. Two kinds of widely used algorithms can be applied to this case: (1) the EM algorithm due to Dempster , Laird and Rubin (1977) and (2) quasi-Newton algorithms (e.g., Nash 1990, and Dennis and Schanbel 1983). In this section we discuss how to find the estimates for the mixed Poisson regression model with a known number of components by combining both algorithms. We also report the results of a Monte Carlo study which investigates the performance of our codes and some implementation issues which will be discussed later.

2.5.1 EM and Quasi-Newton Algorithms

For a *fixed* number of components c, we obtain maximum likelihood estimates of the parameters in the above model using the EM algorithm (Dempster, Laird and Rubin (1977)). As is now standard in mixture model estimation, we implement it by treating unobservable membership of the observations as missing data and representing a complete data set for the model. We discuss choice of number of components below.

Suppose that $(Y, X^{(r)}, X^{(m)}, T) \equiv \{(y_i, x_i^{(r)}, x_i^{(m)}, t_i); i = 1, \ldots, n\}$ is the observed data generated by the mixed Poisson regression model. Let $(Y, Z, X^{(r)}, X^{(m)}, T) \equiv \{(y_i, z_i, x_i^{(r)}, x_i^{(m)}, t_i); i = 1, \ldots, n\}$ be the complete data for the mixture, where the unobserved quantity $z_i = (z_{i1}, \ldots, z_{ic})'$ satisfies

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

The log likelihood for the complete data is

$$l^c(\alpha, \beta \mid Y, Z, X, T) = \sum_{i=1}^{n} \sum_{j=1}^{c} z_{ij} \log(p_{ij}) + \sum_{i=1}^{n} \sum_{j=1}^{c} z_{ij} \log \mathrm{Po}\,(y_i \mid \lambda_{ij}),$$

where p_{ij} and $\mathrm{Po}\,(y_i \mid \lambda_{ij})$ are given by (2.7), (2.8) and (2.9) 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. At the M-step, it finds the parameter estimates which maximize the expected log likelihood for the complete data, conditional on the expected values of the missing data. In our case, this procedure can be stated as follows.

E-step: Given $\alpha^{(0)}$ and $\beta^{(0)}$, replace the missing data Z by its expectation conditioned on these initial values of the parameters and the observed data, $(Y, X^{(r)}, X^{(m)}, T)$. 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 $\tilde{z}_{i,j}(\alpha^{(0)}, \beta^{(0)})$. Then

$$\begin{aligned}
\tilde{z}_{i,j}(\alpha^{(0)}, \beta^{(0)}) &= E\left(z_{ij} \mid \alpha^{(0)}, \beta^{(0)}, Y, M, X^{(m)}, X^{(r)}\right) \\
&= Pr\left(z_{ij} = 1 \mid \alpha^{(0)}, \beta^{(0)}, Y, M, X^{(m)}, X^{(r)}\right) \\
&= \frac{p_j\left(x_i^{(m)}, \beta^{(0)}\right) f_j\left(y_i \mid x_i^{(r)}, t_i, \alpha_j^{(0)}\right)}{\sum_{l=1}^{c} p_l\left(x_i^{(m)}, \beta^{(0)}\right) f_l\left(y_i \mid x_i^{(r)}, t_i, \alpha_l^{(0)}\right)}, \quad j = 1, \ldots, c. \quad (2.20)
\end{aligned}$$

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

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

where

$$\begin{aligned}
Q_1 &= \sum_{i=1}^{n} \tilde{z}_i(\alpha^{(0)}, \beta^{(0)}) \log(p_{ij}) \quad \text{and} \\
Q_2 &= \sum_{i=1}^{n} \tilde{z}_i(\alpha^{(0)}, \beta^{(0)}) \log(\mathrm{Po}(y_i \mid \lambda_{ij})).
\end{aligned}$$

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

$$\frac{\partial Q}{\partial \alpha}\Big|_{\hat{\alpha},\hat{\beta}} = \frac{\partial Q_2}{\partial \alpha}\Big|_{\hat{\alpha}} = \sum_{i=1}^{n} \tilde{z}_i \frac{\partial}{\partial \alpha}[\log(\mathrm{Po}(y_i \mid \lambda_{ij})] = 0 \quad (2.21)$$

$$\frac{\partial Q}{\partial \beta}\Big|_{\hat{\alpha},\hat{\beta}} = \frac{\partial Q_1}{\partial \beta}\Big|_{\hat{\beta}} = \sum_{i=1}^{n} \tilde{z}_i \frac{\partial}{\partial \beta}[\log(p_{ij})] = 0. \quad (2.22)$$

Since closed form solutions of these equations are unavailable, we use a quasi-Newton approach (Nash, 1990) to obtain estimates. This approach makes use of functions Q, and its gradient $g = (\frac{\partial Q}{\partial \alpha}, \frac{\partial Q}{\partial \beta})'$ to find the estimates through an iterative formula

$$(\tilde{\alpha}, \tilde{\beta}) = (\alpha, \beta) + kBg \quad (2.23)$$

where B is a transformation matrix evaluated at (α, β), and k the step length. Note that when B in the above iterative equation equals the inverse Hessian matrix of function Q, this is Newton's method. 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 $\beta^{(0)} = (\beta_1^{(0)}, \ldots, \beta_c^{(0)})$, and two tolerance ϵ_0 and ϵ;

Step 1: (E-step) Compute $\tilde{z}_i = (\tilde{z}_{i,1}, \ldots, \tilde{z}_{i,c})'$, $(i = 1, \ldots, n)$, using (2.20). To avoid overflow in the calculation of $\tilde{z}_{i,j}$, we divide both the numerator and denominator in (2.20) by the largest term in the sum in the denominator;

Step 2: (M-step) Find values of $\hat{\alpha}$ and $\hat{\beta}$ to solve (2.21) and (2.22) using the quasi-Newton algorithm (Nash, 1990). This algorithm consists of two parts: a matrix updating formula for B and a linear search procedure for k in (2.23). Given B and w $(0 < w < 1)$, it chooses $k = 1, w, w^2, \ldots,$ successively until

$$0 < \epsilon_1 < [Q(\tilde{\alpha}, \tilde{\beta}) - Q(\alpha, \beta)]/t^T g \qquad \text{for } \epsilon_1 << 1$$

where $t = (\tilde{\alpha}, \tilde{\beta}) - (\alpha, \beta) = -kBg$ and ϵ_1 is given. Given t, B is updated by

$$\tilde{B} = dtt^T - [t(B\delta g) + (B\delta g)t^T]/t^T\delta g$$

where $\delta g = g(\tilde{\alpha}, \tilde{\beta}) - g(\alpha, \beta)$ and $d = (1 + \delta g^T B \delta g)/t^T \delta g$. Initially, B is set equal to an identity matrix I. Reset $B = I$ if any of the following occurs:

$$\begin{aligned} &\text{(a):} \quad t^T g \leq 0; \\ &\text{(b):} \quad (\tilde{\alpha}, \tilde{\beta}) = (\alpha, \beta); \\ &\text{(c):} \quad t^T \delta g \geq 0. \end{aligned}$$

The stopping criterion for the iterations is

$$\| (\tilde{\alpha}, \tilde{\beta}) - (\alpha, \beta) \| \equiv \sum_{j=1}^{c} \sum_{l=1}^{k_1} | \tilde{\alpha}_{j,l} - \alpha_{j,l} | + \sum_{j=1}^{c} \sum_{l=1}^{k_2} | \tilde{\beta}_{j,l} - \beta_{j,l} | < \epsilon_2$$

where ϵ_2 is a very small positive number;

Step 3: 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, stop.

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

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

(3) $| l(\hat{\alpha}, \hat{\beta} \mid Y, X^{(r)}, X^{(m)}, T) - l(\alpha^{(0)}, \beta^{(0)} \mid Y, X^{(r)}, X^{(m)}, T) | \geq \epsilon_0$, where $l(\alpha, \beta \mid Y, X^{(r)}, X^{(m)}, T)$ is the observed log likelihood function.

Note that we could have used other versions of quasi Newton which use different updating scheme for B.

Dempster, Laird and Rubin (1977) and Wu (1983) discussed the convergence properties of the EM algorithm in a general setting. Since $Q(\alpha, \beta \mid \alpha^{(0)}, \beta^{(0)})$ and its first order partial derivatives are continuous in α, β, $\alpha^{(0)}$ and $\beta^{(0)}$, applying Wu's theorems (1983) in our case, we conclude that the sequence of the observed data likelihood $l(\alpha^{(p)}, \beta^{(p)} \mid Y, X^{(r)}, X^{(m)}, T)$ converges to a local maximum value $l(\alpha^*, \beta^* \mid Y, X^{(r)}, X^{(m)}, T)$, provided that it is not trapped at any saddle point. Furthermore, if $\| \alpha^{(p+1)} - \alpha^{(p)} \| \rightarrow 0$, $\| \beta^{(p+1)} - \beta^{(p)} \| \rightarrow 0$ and the set of local maxima with a given l value is discrete, then $(\alpha^{(p)}, \beta^{(p)})$ converges to (α^*, β^*). Note that for some starting values the stopping criteria in Step 3 above might not be valid. Also $l(\alpha, \beta \mid Y, X^{(r)}, X^{(m)}, T)$ need not, in general, be globally concave. For these reasons, we need to choose initial values carefully in order to increase the chance that the algorithm converges to the global maximum. We will discuss our starting value approach latter.

Note that the above EM algorithm does not directly yield the estimates of the standard errors corresponding to the parameter estimates. On the other hand, when the number of components c is known, asymptotic normality of $\sqrt{n}((\hat{\alpha}, \hat{\beta}) - (\alpha, \beta))$ is easily proved under standard regularity conditions (Lehmann, 1983). To approximate standard error, we compute $\hat{\sigma}(\hat{\alpha}_{j,l})$ and $\hat{\sigma}(\hat{\beta}_{j,l})$ from the diagonal elements of the inverse of the $(c * k_1 + (c-1) * k_2)$-dimensional observed information matrix with c fixed at $\hat{c}$ which is defined as

$$\frac{\partial^2 l(\alpha, \beta \mid Y, X^{(r)}, X^{(m)}, T)}{\partial(\alpha, \beta)^2} = \begin{pmatrix} \frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial \beta} \\ \\ \frac{\partial^2 l}{\partial \alpha \partial \beta} & \frac{\partial^2 l}{\partial \beta^2} \end{pmatrix}.$$

An alternative algorithm to the EM which maximizes the observed log-likelihood $l(\alpha, \beta) \equiv l(\alpha, \beta \mid Y, X^{(r)}, X^{(m)}, T) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{c} p_{ij} \mathrm{Po}(y_i \mid \lambda_{ij})\}$ is a quasi-Newton algorithm (e.g., Nash 1990). Instead of using the E and M steps, we maximize $l(\alpha, \beta)$ by computing successive parameter iterates via the formula

$$(\alpha^{(p+1)}, \beta^{(p+1)}) = (\alpha^{(p)}, \beta^{(p)}) + k B_l g_l$$

where B_l is the transform matrix evaluated at $(\alpha^{(p)}, \beta^{(p)})$, g_l the gradient of $l(\alpha, \beta)$ at $(\alpha^{(p)}, \beta^{(p)})$, and k is a search step length. Note that the maximization of $l(\alpha, \beta)$ is different from maximizing the complete data log-likelihood $Q(\alpha, \beta)$, though the quasi-Newton algorithm is applied in both cases.

In principle either the EM or the quasi-likelihood algorithm can be used to produce the maximum likelihood estimates for the mixed Poisson regression model. The EM and quasi-Newton algorithms, however, have complementary strengths. The convergence rate of the EM algorithm is linear which can be quite slow. In fact adjectives such as 'exceedingly' McCullagh and Nelder (1989), 'maddeningly' Redner and Walker (1984), and 'painfully' Haberman (1977) have been used. As proven by Wu(1983), however, the

EM algorithm converges to a stationary point regardless of the initial guess. A quasi-Newton algorithm on the other hand, often requires rather good initial guesses in order to converge, but the convergence rate in a neighborhood of the solution is much faster than for the EM. The rate is quadratic for a quasi-Newton algorithm.

A sensible combination of these two algorithms is to use the EM until the iterates are in a neighborhood of the solution and finish up with the quasi-Newton algorithm. This is an obvious algorithm to propose and suggestions similar to this have been made. Bock and Aitkin (1981) suggest performing a few EM steps and then one Newton-Raphson step. Dempster et al (1977) suggest using a Newton step while Redner and Walker (1984) suggest switching to a quasi-Newton procedure at some point. Note that using the quasi-Newton algorithm, we can obtain the approximate standard errors of the estimates as by-product.

To combine the EM and the quasi-Newton algorithm for our case, we modify the above Step 3 as follows:

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

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

(3) $| l(\hat{\alpha}, \hat{\beta} \mid Y, X^{(r)}, X^{(m)}, T) - l(\alpha^{(0)}, \beta^{(0)} \mid Y, X^{(r)}, X^{(m)}, T) | \geq \epsilon_0$.

(b) Maximize the observed 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.

2.5.2 Starting Values

We assume that c is known. The first step of our approach divides the data, $\{y_1, \ldots, y_n\}$, into c groups in terms of its percentiles and fits the data into a c-component independent Poisson mixture model without covariates by choosing initial values based on the

percentile information. The second step, if necessary, fits the data into a mixed Poisson regression model containing only one covariate in either Poisson rate or mixing probabilities in such a way that the initial values of the parameters included in the previous mixture model equal the estimates of the corresponding parameters from the previous fitting model, and initial values of the parameters *not* in the previous fitting model are set to a small value, say, 0.00001. This process is iterated until a complete set of initial values for the mixture model is obtained. The motivation of this ad hoc approach is based on the idea of cluster analysis. At each iteration, we use different criteria to classify the data. First, the data are classified in terms of its percentiles. Then the data are classified in terms of independent Poisson mixture model, and subsequently in terms of mixed Poisson regression models. Note that choosing a complete set of initial values for a mixture model step by step in such a way guarantees that the likelihood values will increase in each step. Also our approach obtains maximum likelihood estimates for a sequence of nested mixture models.

We use an example to explain this approach. Suppose that we need to choose initial values to fit a 3-component mixture model with covariates $x_i^{(r)} = (1, d_i)$ and $x_i^{(m)} = (1, e_i)$ where d_i and e_i are real numbers. First, we find 16.5, 33.0, 49.5, 66.0 and 82.5 percentiles of observations $\{y_1, \ldots, y_n\}$ denoted as q_1-q_5 respectively, and fit the data into a 3-component independent Poisson mixture model ($x_i^{(r)} = x_i^{(m)} = (1)$) with the initial values of $\alpha_{1,1}$, $\alpha_{2,1}$ and $\alpha_{3,1}$ equal to $\log(q_1)$, $\log(q_3)$ and $\log(q_5)$ respectively, and both the initial values of $\beta_{1,1}$ and $\beta_{2,1}$ equal to 0. Note that under this specification and an exponential link function, the initial values of $\lambda_j(x_i^{(r)}, \alpha_j)$, $(j = 1, 2, 3)$ are equal to q_1, q_3 and q_5 with the same mixing probabilities 1/3. Second, we fit the data into the 3-component Poisson mixture model with $x_i^{(r)} = (1, d_i)$ and $x_i^{(m)} = (1)$ by choosing the initial values of $\alpha_{1,2}$, $\alpha_{2,2}$ and $\alpha_{3,2}$ equal to 0.00001 and the initial values of the other parameters equal to the estimates of the corresponding parameters of the first fitting

model. Finally, we choose initial values for the 3-component Poisson mixture model with $x_i^{(r)} = (1, d_i)$ and $x_i^{(m)} = (1, e_i)$ in such a way that $\beta_{1,2}$ and $\beta_{2,2}$ are equal to 0.00001 and the other parameters is equal to the estimates of corresponding parameters of the second fitting model.

2.6 A Monte Carlo Study

This section consists of two parts. In the first part, we use Monte Carlo methods to examine the performance of the above algorithm. In particular, we wished to verify the reliability of our code, determine the precision of estimates and investigate some model selection criteria. We use three 3-component mixture models. For each, we analyzed 100 replicates, each with 100 observations. In the second part, we use Monte Carlo methods to study how the mixed Poisson regression models can be used to analyze some typical problems in practice. We also fit the simulated data to Poisson regression models and compare them with the mixed Poisson regression models.

2.6.1 Performance of the Estimation Algorithm

Two different approaches for choosing initial values are compared in the study. In one, we use the true parameter values of the model generating the observations as initial values in order to determine performance of the algorithm in the best case. The other uses the true parameter values of $\alpha_{1,1}$, $\alpha_{2,1}$ and $\alpha_{3,1}$ as initial values, chooses initial values of $\beta_{1,1}$ and $\beta_{2,1}$ according to the approach described in section 2.5.2, and fits the samples to a 3-component independent Poisson mixture model. Then, following the approach of section 2.5.2, we choose a complete set of initial values for the parameters of the model generating the samples. These two different approaches of choosing initial values lead to essentially the same estimates. We describe the details below.

Model 1: A model with Poisson rates depending on one time-dependent covariate, with constant mixing probabilities and $t_i = 1$. For the regression part,

$$x_i^{(r)} = (1,\ d_i), \tag{2.24}$$

where $d_i = 0.2$ for $i = 1, \ldots, 10$, $d_i = 0.4$ for $i = 11, \ldots, 20$, etc., and

$$\alpha \equiv (\alpha_1,\ \ \alpha_2,\ \ \alpha_3) \tag{2.25}$$

where $\alpha_1' = (2.8,\ 2.9)$, $\alpha_2' = (2.6,\ 0.4)$ and $\alpha_3' = (3.6,\ 0.2)$. For the mixing part,

$$\begin{aligned} x_i^{(m)} &= 1 \\ \beta &\equiv (\beta_1,\ \ \beta_2) = (1.1,\ \ 0.6). \end{aligned}$$

For the Poisson rates, we choose an exponential link function defined by

$$\lambda_1(x_i^{(r)}, \alpha_1) = \exp(2.8 - 2.9d_i) \tag{2.26}$$

$$\lambda_2(x_i^{(r)}, \alpha_2) = \exp(2.6 + 0.4d_i) \tag{2.27}$$

$$\lambda_3(x_i^{(r)}, \alpha_3) = \exp(3.6 + 0.2d_i), \tag{2.28}$$

and the mixing probabilities

$$\begin{aligned} p_1(x_i^{(m)}, \beta) &\equiv 0.5156, \\ p_2(x_i^{(m)}, \beta) &\equiv 0.3127 \\ \text{and}\quad p_3(x_i^{(m)}, \beta) &\equiv 0.1717. \end{aligned}$$

Model 2: A model with constant Poisson rates and mixing probabilities depending on one time-dependent covariate. That is, for the regression part,

$$\begin{aligned} x_i^{(r)} &= 1 \\ \alpha &\equiv (\alpha_1,\ \ \alpha_2,\ \ \alpha_3) = (0.4,\ \ 3.0,\ \ 2.0) \end{aligned}$$

and for the mixing part,

$$x_i^{(m)} = (1, \quad d_i) \tag{2.29}$$

where d_i is defined as above, and

$$\beta \equiv (\beta_1, \quad \beta_2) \tag{2.30}$$

where $\beta_1' = (2.0, \ -1.4)$ and $\beta_2' = (-2.0, \ 1.5)$. The Poisson rates, then, are

$$\begin{aligned} \lambda_1(x_i^{(r)}, \alpha_1) &\equiv 1.49, \\ \lambda_2(x_i^{(r)}, \alpha_2) &\equiv 20.08 \\ \text{and} \quad \lambda_3(x_i^{(r)}, \alpha_3) &\equiv 7.39, \end{aligned}$$

and the mixing probabilities are given by

$$p_1(x_i^{(m)}, \beta) = \frac{\exp(2.0 - 2.0d_i)}{\exp(2.0 - 2.0d_i) + \exp(-1.4 + 1.5d_i) + 1} \tag{2.31}$$

$$p_2(x_i^{(m)}, \beta) = \frac{\exp(-1.4 + 1.5d_i)}{\exp(2.0 - 2.0d_i) + \exp(-1.4 + 1.5d_i) + 1} \tag{2.32}$$

$$p_3(x_i^{(m)}, \beta) = \frac{1}{\exp(2.0 - 2.0d_i) + \exp(-1.4 + 1.5d_i) + 1}. \tag{2.33}$$

Model 3: Both the Poisson rates and mixing probabilities depend on the covariate d_i. For the regression part, $x_i^{(r)}$, α and $\lambda_j(x_i^{(r)}, \alpha_j)$ are given by (2.24), (2.25), (2.26), (2.27), and (2.28) respectively; For the mixing part, $x_i^{(m)}$, β and $p_j(x_i^{(m)}, \beta)$ are given by (2.29), (2.30), (2.31), (2.32) and (2.33) respectively.

We chose the above parameter values so that the Poisson rate functions do not cross each other and the ranges of the mixing probabilities for each component do not overlap. We would expect that in this case, the algorithm would perform well.

We carried out these simulations, each with 100 replicates. In each case, the response y_i were obtained by first generating a uniform $(0,1)$ random number u_i and then assigning $y_i \sim \text{Poisson}(\lambda_1(x_i^{(r)}, \alpha_1)$ if $u_i \leq p_1(x_i^{(m)}, \beta)$, $y_i \sim \text{Poisson}(\lambda_2(x_i^{(r)}, \alpha_2))$ if $p_1(x_i^{(m)}, \beta) < u_i \leq p_1(x_i^{(m)}, \beta) + p_2(x_i^{(m)}, \beta)$, or $y_i \sim \text{Poisson}(\lambda_3(x_i^{(r)}, \alpha_3))$ if $u_i > p_1(x_i^{(m)}, \beta) + p_2(x_i^{(m)}, \beta)$.

The results of the Monte Carlo study are presented in Table 2.1. The table shows that for each parameter the mean of estimates is very close to the true value in the models, suggesting that the global maximum of the observed likelihood is reached. For model 1, the sample means are quite close to the true values and the standard deviations are relatively small. Although the Poisson rates of model 2 are estimated accurately, estimates of mixing probabilities are more variable. This suggests that estimating mixing probability parameters in this model is intrinsically more difficult than estimating Poisson rates. This agrees with observations in the literature (Titterington et al., 1985; Mclachlan and Basford, 1988). Estimates of the parameters of model 3 illustrate the same pattern as in Model 2 where estimates of the mixing probability parameters are more variable than those of Poisson rate parameters. Note, however, that although the estimates of mixing probability parameters, $\hat{\beta}$, vary somewhat, the estimated mixing probabilities, $p_j(x_i^{(m)}, \hat{\beta})$, are more precise due to the multimonial link function between the parameters and mixing probabilities.

Our implementation of the algorithm used FORTRAN on a Sun SPARC station 1^+. The average number of the iterations of the EM algorithm for Model 1 is 4.75, 4.93 for Model 2 and 55.6 for Model 3 under the stopping criterion $\epsilon = 0.01$, and average time is 6.65, 7.39 and 79.2 seconds respectively.

2.6.2 The mixed Poisson regression Models For Some Typical Problems

In a clinical trial it may not be uncommon for a treatment to have a significant effect on some subjects but not on others. Thus subjects under treatment may be classified

into two groups: responding and non-responding. Models which ignore this distinction often are unable to detect such a treatment effect. For example, in a clinical trial carried out at British Columbia Children's Hospital which investigated the effect of intravenous gammaglobulin (IVIG) on suppression of epileptic seizures, the clinical investigators conducting this study found that some patients responded to the treatment and others did not. Using Poisson regression to analyze the seizure count data, we found that the data are seriously overdispersed. To explore whether the proposed mixed Poisson regression models can be used to describe and analyze such a scenario, we carried out he following Monte Carlo study.

In the study, we used eight 2-component mixed Poisson regression models in which the mixing probabilities are constant p_1 and p_2, and the Poisson rates are defined by

$$\begin{aligned} \lambda_1(x_i, \alpha_1, \alpha_2) &= \exp(\alpha_1 + \alpha_2 x_i) \\ \text{and} \quad \lambda_2(x_i, \alpha_1) &= \exp(\alpha_1), \end{aligned}$$

where $x_i = 1$ if $i \leq 50$; and 0 otherwise, and $i = 1, \ldots, 100$. This model describes the following situation: there are 50 subjects in each of two groups (e.g., treatment and control groups) for a study which records the observed responses for all subjects; the background effects are characterized by the Poisson rate $\exp(\alpha_1)$; p_1 100% of subjects in the treatment group respond to the treatment which has an effects characterized by the Poisson rate $\exp(\alpha_1 + \alpha_2)$ where $\alpha_2 < 0$; and the other p_2 100% subjects in the treatment group do not respond the treatment, and their responses are the same as the background effects. These eight models in the study are defined by choosing all combinations of parameter values from the following: $p_1 = 0.6$, 0.4, $\alpha_1 = 1.0$, 2.0, and $\alpha_2 = -0.5$, -2.5. Note that the actual Poisson rates of the background effects are 2.7183 and 7.3891 evaluated by $\exp(\alpha_1)$ respectively, and the rates of the treatment effects 1.6487, 4.4817, 0.2231 and 0.6065 by $\exp(\alpha_1 + \alpha_2)$ respectively.

We carried out these simulations, each with 200 replicates. The responses y_i were obtained by first generating a uniform (0,1) random number u_i and then assigning $y_i \sim$ Poisson$(\lambda_1(x_i, \alpha_1, \alpha_2))$ if $u_i \leq p_1$ and $y_i \sim$ Poisson$(\lambda_2(x_i, \alpha_1))$ otherwise. Our implementation of the algorithm used FORTRAN version on a Sun SPARC station 1^+.

The results are reported in Table 2.2 and Table 2.3. It summarizes the properties of the estimated coefficients. Among all eight models the means of $\hat{\alpha}_1, \hat{\alpha}_2$ and $\hat{p}_1$ are very close to the their true values, and their sample standard deviations are very small compared with the magnitudes of the estimates. This means the maximum likelihood estimates are achievable and robust for not only different choices of background and treatment effect but also different choices of responding rates. Since the means and medians of the parameter estimates are very close and upper and lower quartiles are roughly symmetric at the center of the means, the parameter estimates follow approximately normal distributions. Indeed, the histograms of the parameter estimates (not given here) show normal distribution patterns.

To investigate the treatment effect, we test the hypothesis of $\alpha_2 = 0$ by computing the likelihood ratio test statistic. Note that the chi-squared approximation for the likelihood ratio test statistic may not be justified here because the regularity conditions may be not satisfied on the boundary. We use it in these cases as a guideline. The test results are summarized in Table 2.4 in which the numerator in each cell is the number of the times that we reject the hypothesis at 5% significance level, and the denominator is the total number of the replicates. Clearly when the treatment effect is highly significant ($\alpha_2 = -2.5$), we reject the hypothesis of $\alpha_2 = 0$ for almost all replicates at 5 % significance level. This means the likelihood ratio test may work well in these cases. On the other hand, when the treatment effect is small ($\alpha_2 = -0.5$), the likelihood ratio test may not be appropriate partially because the difference between the background and treatment effects may not be significant enough for the test. The baseline effects may not affect the

tests significantly, while the mixing probabilities (respond rate) have some impact on the tests. Note that when $p_1 = 0.4$, there may be only 20 subjects out of 200 who may have a significant treatment effect.

In order to compare the mixed Poisson regression model with Poisson regression, we fitted the simulated data with the Poisson regression model with covariate $(1, x_i)$. The results are summarized in Table 2.5. The means of the intercept estimates in the Poisson regression are very close to the true values in these cases, suggesting that the background effects are appropriately estimated. However the treatment effects are seriously underestimated in these cases because the model cannot distinguish the non-responding subjects from the responding subjects. For example, in the two cases of the low treatment ($\alpha_2 = -0.5$) and low background ($\alpha_1 = 1.0$) effect, the estimate of the treatment effect by the Poisson regression is -0.2668 for the mixing probability $p_1 = 0.6$ and -0.1611 for $p_1 = 0.4$, which are about one half and one quarter of the true parameter value respectively; In the two cases of the high treatment ($\alpha_2 = -2.5$) and high background ($\alpha_1 = 2.0$) effect, the estimate of the treatment effect by the Poisson regression is -0.8065 for the mixing probability $p_1 = 0.6$ and -0.4536 for $p_1 = 0.4$, which are less than one quarter and one fifth of the true value respectively. We also carried out the test for the hypothesis of $\alpha_2 = 0$ using the likelihood ratio test statistic. The test results given in Table 2.6 in which the numerator in each cell is the number of times that we reject the hypothesis, and the denominator is the total number of these tests. For example, in the two cases of the low background ($\alpha_1 = 1.0$) and low treatment ($\alpha_2 = -0.5$) effect, 99 times out of 200 for mixing probability $p_1 = 0.6$ and 47 times out of 200 for $p_1 = 0.4$, respectively, that we reject the hypothesis at 5% significance level; In the two cases of the high background ($\alpha_1 = 2.0$) and high treatment ($\alpha_2 = -2.5$) effect, we *always* reject the hypothesis at 5% significance level for both the mixing probability values. Note that the Poisson regression is more powerful except one case, although Table 2.4 and Table 2.6

have a similar pattern.

Using the mixed Poisson regression model, we can classify subjects as responding and non-responding. In the Monte Carlo study, for $x_i = 1$, y_i is identified with group one generated by Poisson rate $\lambda_1(x_i, \alpha_1, \alpha_2)$ if the estimated posterior probability of being group one $\hat{z}_{i,1} > 0.5$, and with the other generated by Poisson rate $\lambda_2(x_i, \alpha_1)$ otherwise. For 200 replicates the mean of the number of subjects in the treatment group who responded to the treatment is very close to $50p_1$, suggesting that the classification criterion works well.

2.7 Implementation Issues

2.7.1 Model Selection

We need to address the following two issues when applying a mixed Poisson regression model: (a) We must determine the number of components c, and (b) we must have a method to carry out inference about model parameters. When c is known, inference for the parameters can be based on a likelihood ratio test. In practice, however, this is rarely the case. When c is unknown, the likelihood ratio test is no longer valid for determining c or testing hypotheses about parameter values. This is because the usual regularity conditions do not hold for the likelihood ratio test statistic to have its standard asymptotic null distribution of chi-squared with degree of freedom equal to the difference between the number of parameters under the null and alternative hypotheses. One of the regularity conditions requires that the parameters in a mixture are identifiable without any restriction. This ensures that the information matrix is non-singular. The main problem here is the lack of identifiability even when the class of the mixed Poisson regression models is identifiable. As McLachlan and Basford (1988) illustrate this, consider a 2-component

mixture without covariates. The null hypothesis that there is one underlying population,

$$H_0 : c = 1,$$

can be approached by testing whether $p_1 = 1$, which is on the boundary of the parameter space with a consequent breakdown in the standard regularity conditions. Alternatively, we can view H_0 as testing for whether $\lambda_1 = \lambda_2$, where now the value of p_1 is irrelevant. If for a specified value of p_1 regularity conditions held, so that the log likelihood ratio test statistic under H_0 were distributed asymptotically as chi-squared, then the null asymptotic distribution of the likelihood ratio test statistic where p_1 is unspecified, would correspond to the maximum of a set of dependent chi-squared variables. A comprehensive account of the breakdown in regularity conditions has been give by Ghosh and Sen (1985); see also Hartigan (1985a,b), Titterington, Smith and Makov (1985), and Mclachlan and Basford (1988). We propose the following methods for model selection.

In general, there are two criteria used for statistical model selection: the principle of parsimony and closeness to the true distribution. The former means that more parsimonious use of parameters should be pursued so as to raise the accuracy of estimates for unknown parameters in a model. On the other hand, closeness to the true model is incompatible with parsimony of parameters. These two criteria form a trade-off: if one pursues one of the criteria, the other must be necessarily sacrificed. The multiple correlation coefficient adjusted for the degrees of freedom may be most commonly used statistic that incorporates these two incompatible criteria into a single statistic.

Akaike (1973) has proposed a more general as well as more widely applicable statistic that ingeniously incorporates the above two criteria. As it is based on the Kullback-Leibler Information Criterion (KLIC), Akaike's statistic is called Akaike Information Criterion and is abbreviated as the AIC. The AIC can be derived as follows.

Suppose that the adequacy of a postulated model $F(y \mid \theta)$ to approximate the unknown true distribution $G(Y)$ is measured by the KLIC

$$I(G : F(\cdot \mid \theta)) = E_G[\log \frac{g(Y)}{f(Y \mid \theta)}]$$

where θ is a finite-dimensional vector of unknown parameters; g and f are density (or probability) functions of G and F respectively; $E_G(\cdot)$ stands for expectation with respect to the true distribution G. We define a pseudo-true model $F(\cdot \mid \theta_0)$ with a parameter value θ_0 such that

$$I(G : F(\cdot \mid \theta_0)) \leq I(G : F(\cdot \mid \theta))$$

for any possible θ in the admissible parameter space. The model $F(\cdot \mid \theta_0)$ may be regarded as the most adequate relatively within the family models $F(y \mid \theta)$ in the sense that the KLIC for $F(y \mid \theta)$ is minimized by $F(y \mid \theta_0)$.

Assuming that $I(G : F(\cdot \mid \theta_0)) = O(n^{-1})$, i.e., the pseudo-true model is nearly true, Akaike (1973) derives

$$AIC(F(\cdot \mid \theta)) = -2 \log f(y \mid \hat{\theta}) + 2k$$

as an almost unbiased estimate for $-2E_G[\log f(Y \mid \theta_0)]$ where $\hat{\theta}$ is the maximum likelihood estimate for θ based on observation y and k is the number of unknown parameters, i.e., the dimension of θ. Note that the first term of the AIC measures the goodness-of-fit of the model to a given set of data, because $f(y \mid \hat{\theta})$ is the maximized likelihood function. The second term is interpreted as representing a penalty that should be paid for increasing the number of parameters. In this sense the AIC may be regarded as an explicit formation of the so-called principle of parsimony in model building.

Schwartz (1978) has proposed another model selection criterion: the Bayesian Information Criterion (BIC). The BIC is defined through a larger-sample version of Bayes

procedures by placing a prior distribution on the parameter space including all dimensions and models considered. It can be derived as follows.

We assume that observations are generated by a distribution from a family with a density

$$f(y, \theta) = \exp(\theta \cdot x(y) - b(\theta)),$$

where $\theta \in \Theta$, a convex subset of the K-dimensional Euclidean space, and $x(y)$ is the sufficient K-dimensional statistic. The competing models are denoted by sets $m_j \in \Theta$, where m_j is a k_j- dimensional linear submanifold of K-dimensional space.

Since the a priori distribution need not be known exactly for the asymptotic results, we assume that it is of the form $\sum \alpha_j \mu_j$, where α_j is the a priori probability of the jth model being the true one, and μ_j, the conditional a priori distribution of θ given the j model, has a k_j-dimensional density that is bounded and locally bounded away form zero throughout $m_j \in \Theta$.

Finally, we assume a fixed penalty for guessing the wrong model. Under this assumption, the Bayes solution consists of selecting the model that is a posterior most probable. That is equivalent to choosing the j that maximizes

$$S(X, n, j) = \log \int \alpha_j \exp(X \cdot \theta - b(\theta)n) d\mu_j(\theta),$$

where the integral extend over $m_j \in \Theta$, and X is the averaged x-statistic $(1/n) \sum X(y_i)$.

For fixed X and j, as n tends to infinity, we obtain the asymptotic expansion of $S(X, n, j)$ as

$$S(X, n, j) = n \sup(X \cdot \theta - b(\theta)) - \frac{1}{2} k_j \log n + R,$$

where the remainder $R = R(X, n, j)$ is bounded in n for fixed x and j. Therefore, for a large sample, maximizing $S(X, n, j)$ in j is equivalent to maximizing

$$BIC = \log f_j(y_1, \ldots, y_n) - \frac{1}{2} k_j \log n,$$

where $f_j(y_1, \ldots, y_n)$ is the maximum likelihood function for model j, and k_j is the dimension of the model.

Qualitatively both the AIC and BIC give a mathematical formulation of the principle of parsimony in model building. Quantitatively, since the BIC differs from the AIC only in that the dimension is multiplied by $(\log n)/2$, the BIC leans more than the AIC towards lower-dimensional models. For large numbers of observations the two model selection procedures differ markedly from each other.

McLachlan and Basford (1988) discussed the use of AIC to determine the number of components in a finite mixture model. Leroux and Puterman (1992) applied AIC and BIC to select independent Poisson mixture models. We define the AIC and BIC criteria for the mixed Poisson regression model as follows:

- AIC: choose the model for which $\hat{l}_c(X) - a_c(X)$ is largest;
- BIC: choose the model for which $\hat{l}_c(X) - \frac{1}{2}(\log(n))a_c(X)$ is largest

where $\hat{l}_c(X)$ is the maximum log-likelihood of the mixture with c components and covariate X, $a_c(X) = c * k_1 + (c-1) * k_2$ where k_1 and k_2 are the dimensions of α_j and β_j respectively, and n is the total number of observations. As discussed above, these two criteria do not always select the same model; the BIC tends to select a smaller number of components than AIC when there are 8 or more observations.

Using the BIC (AIC), our model selection approach consists of two stages. At the first stage, we determine c to maximize BIC (AIC) values for the saturated 1-3 (1-4) component mixture models that contain all possible covariates in both rates and mixing probabilities. Although we compute both AIC and BIC values in our applications, we recommend using BIC because Monte Carlo studies reported below suggest that BIC is more reliable in the model selection. At the second stage, our model selection approach depends on our analysis objectives. If our goal is inference about some particular model

parameter, we carry out likelihood ratio tests for nested c-component mixture models. If the goal is choosing an appropriate model to fit the data, we select a model to maximize BIC (AIC) values among c-component mixture models concerned. Since this selection method is heuristic and only gives a guideline in applications, some other specific concerns in model selection should be taken into account from case to case. For instance, in some applications the number of components and some parameters in a mixture may be explicitly or implicitly determined by underlying theory, especially when a mixture model is intended as a direct representation of the underlying physical phenomenon. For a housing market in disequilibrium, the market has two phases: supply and demand. If we regard the phase in operation in any given month to be the unobservable underlying state because it may not be clear which phase is in operation, we have a two-component mixture model. Goldfeld and Quandt (1973) discuss such a model and denote it as a switching regression model.

In the Monte Carlo studies discussed in Section 2.6.1, we computed both AIC and BIC values for all possible mixed 2 to 4 component models. Table 2.7 shows that AIC and BIC are reliable methods for choosing the correct models. AIC chose the correct model 96% of the time for Model 1, 87% of the time for Model 2 and 91% of the time for Model 3. When AIC failed to select the correct model, it always chose a model with too many components, suggesting that AIC may under-penalize the number of parameters in the mixtures. On the other hand, BIC *always* chose the correct models, suggesting that BIC may not over-penalize the number of parameters. Note that all sample sizes in the Monte Carlo studies are 100. The examples in the next section will exhibit this procedure in practice.

2.7.2 Classification

In classification, the number and composition of groups are not known at the start of the investigation. On occasion, the aim of a classification study may be to enable the subsequent assignment of new objects. For instance, in pattern recognition (Fukunaga, 1972, and Duda and Hart, 1973), information about 'patterns' can be obtained from a 'training' set of observations which may be analyzed by classification method.

Fitting the mixed Poisson regression models to Poisson-distribution data, we assume that each observation belongs one of c groups characterized by the Poisson rate functions. One possible use of the mixed Poisson regression model is to classify data on the basis of a probabilistic model rather than an ad hoc clustering technique. Since $\tilde{z}_{i,j}$ in (2.20) is the estimated posterior probability that the ith observation y_i is generated by the jth component distribution $f_j\left(y_i \mid x_i^{(r)}, t_i, \alpha_j\right)$, this information can be used to classify observations into different groups characterized by the component distributions. For instance, for a c-component mixture model we may postulate c different groups defined by the c different forms of Poisson rates, $\lambda_j(x_i^{(r)}, \alpha_j)$ $(j = 1, \ldots, c)$ of the model. According to the classification criterion, an observation i is identified with the component which maximizes $\tilde{z}_{i,j}$. In our Monte Carlo study this classification criterion works very well. Also in our applications, maximum values for this quantity all exceed 0.5. Note that if the parameters of the model were known, this classification criterion would be the optimal or Bayes rule (Anderson, 1984, chapter 6) which minimizes the overall error rate. Also such a approach has been referred to as latent class analysis (Aitkin et al. 1981). We illustrate this approach in examples below.

2.7.3 Residual Analysis and Goodness-of-fit Test

Once a mixed Poisson regression model has been fit to a set of observations, it is essential to the quality of the fit. For this purpose, we consider Pearson, deviance and likelihood residuals for mixed Poisson regression models, and use them to identify individually poorly fitting observations and influential observations on overall fit of the model as well. We also define a quantity to measure influence of individual observations on the set of parameter estimates, and use it to identify influential observations. In addition, we give goodness-of-fit statistics for mixed Poisson regression models.

Definitions of Residuals

For Normal regression models, we can express an observation y_i of the response variable in the form

$$y_i = \hat{\mu}_i + (y_i - \hat{\mu}_i),$$

where $\hat{\mu}_i$ is the maximum likelihood estimate of the mean of y_i, i.e., data=fitted value +residual. Residuals are used in many procedures designed to detect various types of disagreement between data and assumed model. For example, the scatterplot of residuals versus fitted values that accompanies a linear least square fit is a standard tool used to diagnose nonconstant variance, curvature, and outliers. Diagnostic tools such as this plot have two important uses. First, they may result in the recognition of important phenomena that might otherwise have gone unnoticed. Outlier detection is an example of this, where an outlying case may indicate conditions under which a process works differently, possible worse or better. Second, the diagnostic methods can be used to suggest appropriate remedial action to the analysis of the model.

For generalized linear models there are at least three types of generalized residuals

which are widely used in practice. One is the *Pearson residual* defined as

$$r_{pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}, \tag{2.34}$$

where $V(\mu)$ is the variance function and $\hat{\mu}_i$ is the maximum likelihood estimate of the ith mean of fitted to the regression model. These residuals are the signed square roots of the contribution to the Pearson goodness-of-fit statistic X^2. For the usual Poisson regression model, the Pearson residual is

$$r_{pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}},$$

where $\hat{\mu}_i = \exp(x_i'\hat{\alpha})$ and $\hat{\alpha}$ is the maximum likelihood estimates of the regression parameters; for the usual logistic regression model,

$$r_{pi} = \frac{y_i - m_i\hat{p}_i}{\sqrt{m_i\hat{p}_i(1-\hat{p}_i)}},$$

where $\hat{p}_i = logit(x_i'\hat{\alpha})$.

The second type of generalized residuals is deviance residual defined as

$$\begin{aligned} r_{di} &= \text{sign}(y_i - \hat{\mu}_i)\sqrt{2[l(y_i, y_i) - l(\hat{\mu}_i, y_i)]} \\ &= \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \end{aligned} \tag{2.35}$$

where $l(\mu_i, y_i)$ is the log likelihood function for y_i and d_i is the contribution to the deviance goodness-of-fit statistic D. For the usual Poisson regression model,

$$d_i = 2(y_i \ln(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i),$$

where $\hat{\mu}_i = \exp(x_i'\hat{\alpha})$; for the usual logistic regression model,

$$d_i = 2y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + 2(m_i - y_i)\ln\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right),$$

where $\hat{\mu}_i = m_i\hat{p}_i = m_i logit(x_i'\hat{\alpha})$. The third type of generalized residuals is the likelihood residual which is derived by comparing the deviance obtained on fitting a linear 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 data set. The likelihood residual for the ith case is defined as

$$r_{li} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{D - D_{(i)}} \tag{2.36}$$

where D and $D_{(i)}$ are the deviances based on n and $n-1$ cases respectively. Pregibon (1981) derives useful one step approximation for the above exact value by

$$r_{li} \simeq \text{sign}(y_i - \hat{\mu}_i)\sqrt{\frac{h_i}{1-h_i} r_{pi}^2 + r_{di}^2}$$

where h_i is the ith diagonal element of the $n \times n$ matrix

$$H = W^{1/2} X (X'WX)^{-1} X' W^{1/2}. \tag{2.37}$$

In this expression for H, W is the $n \times n$ diagonal matrix of weights used in fitting the linear model and X is the $n \times k$ design matrix. For Poisson regression model, $W = \text{diag}\{\hat{\mu}_1, \ldots, \hat{\mu}_n\}$; for logistic regression model, the ith diagonal element of W is $m_i \hat{p}_i (1 - \hat{p}_i)$.

Note that when the response Y_i follows a Normal distribution, r_{li}^2 follows χ_1^2 distribution; when Y_i follows a non-Normal distribution, r_{li}^2 does not asymptotically follows χ_1^2 distribution as $n \to \infty$ because the asymptotical theory does not hold in this case (Williams, 1987).

To standardize the above residuals so that they have approximate unit variance, one needs to account for the inherent variation in the fitted values $\hat{\mu}_i$. In general, for any type of residuals $R(y_i, \hat{\mu}_i)$, Pierce and Schafer (1986) show that its variance is approximately given by

$$Var[R(y_i, \hat{\mu}_i)] \simeq Var[R(y_i, \mu_i)] - Var[(\hat{\mu}_i - \mu_i)/SD_{u_i}(y_i)] \tag{2.38}$$

as $\mu_i \to \infty$. For either Poisson regression or logistic regression model,

$$Var[(\hat{\mu}_i - \mu_i)/SD_{u_i}(y_i)] = h_i$$

where h_i is defined by (2.37). Therefore, we can standardize r_{pi}, r_{di} and r_{li} by dividing the factor $\sqrt{1-h_i}$ because for all three types of the residuals the first term in the right side of (2.38) is 1.

Several researchers have compared differences between these three types of residuals (e.g., Pierce and Schafer, 1986; Williams, 1987; McCulagh and Nelder, 1989; and Collett, 1991). The value of r_{li} is intermediate between r_{di} and r_{pi}, and it is usually much closer to r_{di} than to r_{pi}. Both r_{di} and r_{li} take account of the shape of the distribution of Y_i which is ignored by r_{pi}. Both r_{di} and r_{li} have distributions which are closer to normality than that of r_{pi}. For outlier detection r_{li} seems the best choice because of its relevance to the measurement of case influence on likelihood ratio tests.

Several types of residual plots are useful for different purposes of diagnostics. For example, an index plot that the residuals are displayed against the corresponding observation number or index is particularly suitable for detection of outliers. Although a plot of the residuals against the fitted values $\hat{\mu}_i$ or an explanatory variable is more informative than an index plot for normal regression, it may be uninformative for Poisson regression because when the mean of the response variable is small, there may be a pattern in the plot no matter whether the model is correct or not. Indeed, if $y_i = 0$, $r_{pi} = r_{di} = -\sqrt{\hat{\mu}_i}$. This means that for small mean values, the residuals are not approximately normal.

Analogously, we can define the same three types of residuals for mixed Poisson regression models. That is, the *Pearson residual*, r_{Pi}, for mixed Poisson regression models is given by defining $\hat{\mu}_i$ and $V(\hat{\mu}_i)$ in (2.34) as

$$\hat{\mu}_i = t_i \sum_{j=1}^{c} \hat{p}_{ij}\hat{\lambda}_{ij}, \tag{2.39}$$

where

$$\begin{aligned}
\hat{\lambda}_{ij} &= \exp(\hat{\alpha}_j' x_i^{(r)}), \\
\hat{p}_{ij} &= \frac{\exp(\hat{\beta}_j' x_i^{(m)})}{\sum_k^{c-1} \exp(\hat{\beta}_k' x_i^{(m)}) + 1} \quad \text{for } j = 1, \ldots, c-1 \text{ and} \\
\hat{p}_{ic} &= \frac{1}{\sum_k^{c-1} \exp(\hat{\beta}_k' x_i^{(m)}) + 1},
\end{aligned}$$

and

$$V(\hat{\mu}_i) = t_i \sum_{j=1}^{c} \hat{p}_{ij} \hat{\lambda}_{ij} + t_i^2 \left\{ \sum_{j=1}^{c} \hat{p}_{ij} \hat{\lambda}_{ij}^2 - [\sum_{j=1}^{c} \hat{p}_{ij} \hat{\lambda}_{ij}]^2 \right\}.$$

The deviance residual, r_{Di}, for mixed Poisson regression models is given by defining the log likelihood function $l(\hat{\mu}_i, y_i)$ in (2.35) as

$$l(\hat{\mu}_i, y_i) = \log[f(y_i \mid x_i^{(r)}, x_i^{(m)}, t_i, \hat{\alpha}, \hat{\beta})] \tag{2.40}$$

where $f(y_i \mid x_i^{(r)}, x_i^{(m)}, t_i, \alpha, \beta)$ is defined in (2.10). Note that $l(y_i, y_i)$ is the same for both generalized linear models and mixed Poisson regression models because we have the following relation

$$\begin{aligned}
f(y_i \mid x_i^{(r)}, x_i^{(m)}, t_i, \alpha, \beta) &= \sum_{j=1}^{c} p_{ij} \mathrm{Po}\,(y_i \mid \lambda_{ij}) \\
&\leq \sum_{j=1}^{c} p_{ij} \mathrm{Po}\,(y_i \mid y_i) \\
&= \mathrm{Po}\,(y_i \mid y_i)
\end{aligned}$$

This indicates that there is the same baseline for generalized linear models and mixed Poisson regression models.

The likelihood residual r_{Li} for mixed Poisson regression models is given by defining $\hat{\mu}_i$ as specified in (2.39), and D and $D_{(i)}$ as the deviances based on the data set of n and

$n-1$ cases for the mixed Poisson regression model. Computing the likelihood residuals requres fitting the model n times, each having good starting values which are already available in our algorithm. In contrast to linear normal regression, it may require fitting the model only once.

Note that for the residuals of mixed Poisson regression models, equation (2.38) still hold. Thus, to account for variation in the fitted values $\hat{\mu}_i$ in these three types of the residuals, we need to calculate

$$Var[(\hat{\mu}_i - \mu_i)/SD_{u_i}(y_i)].$$

However, the computation of this variance now becomes too complicated. Fortunately, for large samples, $\hat{\mu}_i$ are very close to μ_i so that the variation in the fitted values may be negligible.

Example. R&D and Patent In modeling the patent data from Section 2.8.1 on the relationship between R&D spending and number of patent applications at firm level, a 3-component mixed Poisson regression model is found to be satisfactory. The analysis will be given in Section 2.8.1. Figure 2.1, Figure 2.2 and Figure 2.3 give index plots of the Pearson, deviance and likelihood residuals respectively.

Figure 2.1 shows that the Pearson residuals may not aprroximately be normal. On the other hand, Figure 2.2 and Figure 2.3 show that the deviance and likelihood residuals are very similar to each other. Note that the 6th has the largest Pearson residual and the 8th has both largest deviance and likelihood residuals. These plots suggest that the deviance residuals and the likelihood residuals may be likely to perform similarly in terms of the ranking of extreme observations. In fact, the empirical evidence to be presented in examples in Section 2.8 suggest the same. The numerical studies also indicate that r_{Di} and r_{Li} are more approximately normal than r_{Pi}. Since the likelihood residuals are much more difficult to compute than any other type of residuals, we recommend using

r_{Di} routinely.

Detection of Outliers and Influential Observations

The residuals obtained after fitting a mixed Poisson regression model to an observed set of data form the basis of diagnostic techniques for assessing model adequacy. Since our primary objective of residual analysis for mixed Poisson regression models is to identify outliers and influential observations, we discuss how these residuals can be used for this objective.

Like generalized linear models, we define outliers as those observations that are surprisingly distant from the remaining observations in the sample. Such observations may occur as a result of measurement errors, that is errors in reading, calculating or recording a numerical value; or they may be just an extreme manifestation of natural variability.

Since large residuals indicate poorly fitting observations, we use index plots of residuals for detection of outliers, that is, observations that have unusually large residuals. For example, in the previous example, the 8th observation stands out from the rest as having a relatively large residual in all three index plots of the residuals. The outlying nature of this observation is obvious from these plots.

The influence of a particular observation on the overall fit of a model can be assessed from the change in the value of a summary measure of goodness of fit that results from excluding the observation from the data set. Since r_{Li}^2 is the change in deviance on omitting the ith observation from the fit, an index plot of these values is the best way of assessing the contribution of each observation to the overall goodness of fit of the model. In the previous example, Figure 2.3 shows that the 8th observation has great impact on the overall fit of the model to the data, as measured by the deviance. Indeed, on omitting the 8th observation, the deviance reduction is $r_{L8}^2 = (3.392)^2 = 11.506$.

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

$$\begin{aligned} w_i &= \frac{1}{p}\left\{\| (\hat{\alpha} - \hat{\alpha}^{(i)})/se(\hat{\alpha}) \| + \| (\hat{\beta} - \hat{\beta}^{(i)})/se(\hat{\beta}) \|\right\} \\ &= \frac{1}{p}\left\{\sum_{j=1}^{c}\sum_{l=1}^{k_1}\frac{|\hat{\alpha}_{j,l} - \hat{\alpha}_{j,l}^{(i)}|}{se(\hat{\alpha}_{j,l})} + \sum_{j=1}^{c-1}\sum_{l=1}^{k_2}\frac{|\hat{\beta}_{j,l} - \hat{\beta}_{j,l}^{(i)}|}{se(\hat{\beta}_{j,l})}\right\} \end{aligned} \tag{2.41}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum parameter estimates of the mixed Poisson regression model based on the complete data set of n cases, and $\hat{\alpha}^{(i)}$ and $\hat{\beta}^{(i)}$ on the data set of $n-1$ cases excluding the i case; $se(\hat{\alpha})$ and $se(\hat{\beta})$ are the estimated standard errors of the corresponding estimates based on the n cases, and $p = ck_1 + (c-1)k_2$. Because each term in (2.41) measures a relative change in individual coefficient, w_i may be interpreted as average relative coefficient changes 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.

For the previous example, Figure 2.4 is the index plot of w_i. Clearly, the plot shows that the 8th, 12th, 47th 64th, 65th and 66th observations are influential so that omitting each of them from the data has a great effect on the set of parameter estimates. For example, if the 12th observation is excluded from the data set, each parameter estimate will averagely change about 33%. Although the 47th observation has relatively large value of w_i, it has a relatively small value of either likelihood residual or deviance residual. This indicates that an influential observation need not necessarily be an outlier. In particular, an influential observation that is not an outlier will occur when the observation distorts the form of the fitted model to such an extent that the observation itself has a small residual value. Note that in this example, the 8th observation is not only an influential

observation but also an outlier as well. On the other hand, the first observation appears an outlier but has a rather small value of w_i.

Goodness-of-fit Statistics

After fitting a mixed Poisson regression model to a set of data, it is natural to inquire about the extent to which the fitted values of the response variable under the model compare with the observed values. If the agreement between the observations and the corresponding fitted values is good, the model may be acceptable. If not, the current form of the model will certainly not be acceptable and the model will need to be revised. The aspect of the adequacy of a model is widely referred to as goodness of fit.

There are at least two widely used goodness-of-fit statistics which can be used here. One is the deviance statistic, D, defined as

$$D = \sum_{i=1}^{n} r_{Di}^2$$

where r_{Di} is the deviance residuals for the mixed Poisson regression model; And the other is the Pearson's statistic, X^2, defined as

$$X^2 = \sum_{i=1}^{n} r_{Pi}^2$$

where r_{Pi} is the Pearson residuals for the mixed Poisson regression model. In order to evaluate the extent to which an adopted mixed Poisson regression model fits a set of data, the distribution of either the deviance or the Pearson statistic, under the assumption that the model is correct, is needed. For normal linear models, the deviance and the Pearson's X^2 statistics are distributed as χ^2 with $(n-p)$ degrees of freedom, where n is the number of observations and p is the number of unknown parameters in the model. In general, many studies have shown that the Pearson statistic is often much more nearly

chi-squared than that of the deviance (e.g., Larntz, 1978). For this reason, we use the Pearson statistic for overall goodness of fit tests for the mixed Poisson regression models.

2.8 Applications

2.8.1 R&D and Patents

Economists studying technological innovation often use patent applications as an indicator of inventive activity. The nature of much industrial R&D activity suggests that it is natural to assume that patent counts follow a Poisson distribution: patent applications can be thought of as measuring the number of successful outcomes among a large (but unobserved) number of projects within a firm's R&D lab, each of which has a small probability of success. Econometricians have accordingly examined the relationship between R&D and patenting by using Poisson regression to estimate a "production function for patents" of the form: $E(y_i) = \exp(\alpha' x_i)$, where y_i is the number of patents applied for by firm i and x_i is a vector of explanatory variables, including R&D spending. (There are many problems with using patent counts as indicators of innovative output, but they remain the only comprehensive, objective, and readily available measure of inventive activity. See Griliches (1990).)

In economics there are two important characteristics associated with a production function $f(x)$: returns to scale and elasticity. The former identifies how output responds to proportionate, scaled expansion in inputs. If a proportionate increase in all inputs increases output by the same proportion, the production function is said to exhibit constant returns to scale. This can be mathematically described by

$$tf(x) = f(tx),$$

where x is a vector representing inputs, and t is a positive real number. Similarly, if a more (less) than proportionate increase (decrease) in output is obtained, there is increasing

(decreasing) returns to scale. These can be mathematically described by

$$tf(x) < f(tx)$$

and

$$tf(x) > f(tx)$$

respectively.

An input (x_i) elasticity of output is a measure of responsiveness of output to that input that uses the percent change in output divided by the percent change in the input. This is given by

$$\epsilon_i = \frac{\Delta f/f}{\Delta x_i/x_i} = \frac{\partial(\log f)}{\partial(\log x_i)}$$

Note that for the above production function for patents, for instance, the R&D elasticity of patent applications is independent of the units in which patents are measured, and thus a more meaningful measure of the responsiveness of patent applications to R&D spending. The R&D elasticity of patent applications simply measures the percentage change in patent applications when R&D spending changes by a small percent.

The parameters of the above model have a direct and interesting economic interpretation: they provide estimates of returns to scale in performing R&D. However, efforts to test for returns to scale using these data have been hampered by the fact that they are typically quite severely overdispersed. Hausman, Hall and Griliches (1984), Bound et al. (1984), and Hall, Griliches, and Hausman (1986) estimated variations of the Poisson model which account for the overdispersion by including an additive random firm effect in the patent equation. The random firm effect can be thought of as capturing unobserved firm-specific factors affecting R&D productivity. As is well known, if the additive random effect is distributed gamma then the unconditional distribution of the response variable

is negative binomial. If the distribution assumption is incorrect, inconsistent parameter estimates will be obtained. These studies also present results from the quasi-generalized pseudo maximum likelihood estimators proposed by Gourieroux, Monfort, and Trognon (1984) which allow the random firm effect to be drawn from an unspecified distribution. Though results obtained using the Poisson, Negative Binomial, and QGPML estimators were qualitatively the same, the estimated coefficient on R&D varied substantially. The authors attributed this problem to "instability" in the R&D-patents relationship over time and across firms.

We treat the unobserved heterogeneity in these data quite differently, and show how the overdispersion can be accounted for in an alternative and perhaps more interesting way by using finite rather than continuous mixtures. Rather than assume that all firms have common regression coefficients and a random intercept, we allow both the intercept and the coefficient on R&D to vary from firm to firm, but in a restricted way. We postulate a *discrete* Poisson mixture model in which firms can be in a finite number of different states defined by different degrees of R&D productivity, for example "high", "medium", and "low". In this model the coefficients vary from state to state, rather than from firm to firm. One way to motivate this model is to assume that all firms have access to the same technological opportunities, but have different unobservable innovative capabilities (e.g. "Type A" or "Type B" or "Type C" organizational structures). Alternatively, we could assume that all firms have the same innovative capabilities, but have differential access to technological opportunities: some firms are working in "hot" areas of the underlying science while others are not.

The data are patent applications and R&D spending in 1976 for 70 pharmaceutical and biomedical companies, taken from the NBER R&D Masterfile (see Hall (1988) for documentation of this data set.) The data are displayed in Figure 2.5, where the horizontal axis is the logarithm of R&D. Formal test results in Table 2.8 confirm the visual

impression that the data are overdispersed: all of the tests strongly reject the null hypothesis of no overdispersion. As in the standard model used in previous studies, the dependent variable is a count of patent applications, and the explanatory variables are $\log(R\&D)$ and a quadratic term $(\log(R\&D))^2$ included to capture non-linearities in the relationship. The coefficients on these variables provide a direct estimate of the elasticity of innovative output with respect to R&D spending, and thus the extent to which there are scale economies in performing R&D. If the elasticity is greater than one then an increase in R&D spending would generate a more than proportionate increase in patents. The coefficient on the quadratic term is particularly interesting since it captures the extent to which economies of scale vary with the size of a firm's R&D effort, a question which has been been hotly (though inconclusively) debated by economists for many years.

To apply our mixture model, we assume that

(1) the total number of patents applied for by firm i is associated with covariates $x_i = (x_i^{(m)}, x_i^{(r)})$, where $t_i = 1$ (one year), $x_i^{(m)} = (1)$ and $x_i^{(r)} = (1, \log(R\&D_i), (\log(R\&D_i))^2)$ where $R\&D_i$ is R&D expenditure of firm i in 1976. Note that $x_i^{(m)} = (1)$ correspond to the assumption of constant mixing probabilities. Note also that the mixing probability here may be interpreted as the likelihood that a firm stays in a particular underlying state during one year period. Since R&D expenditure is usually calculated at the end of a year, for one year patent data, it is legitimate to assume that the mixing probabilities are independent of R&D covariates;

(2) patent counts of different firms are independent;

(3) each patent count follows a mixed Poisson distribution with Poisson rates defined by exponential link functions

$$\lambda_j(x_i^{(r)}, \alpha_j) = \exp[\alpha_{j0} + \alpha_{j1}\log(R\&D_i) + \alpha_{j2}(\log(R\&D_i))^2]$$

where $i = 1, 2, ..., 70$, $j = 1, 2, ..., c$, and c is the number of components in the mixture.

The maximum likelihood estimates for the saturated 1-4 component mixture models and several constrained 3-component mixture models applied to the data are given in Table 2.9. Among the four saturated mixture models, both AIC and BIC lead to the choice of 3-component mixtures. Within the class of 3-component mixture models, the saturated 3-component mixture model is considered as the most appropriate one to fit the data in terms of BIC (AIC).

After fitting the 3-component mixed Poisson regression model to the data, the Pearson goodness-of-fit statistic X^2 is 64.53 with 59 degrees of freedom. This value does not exceed the upper 95% critical point of the χ^2-distribution on 59 degrees of freedom, $\chi^2_{59,0.95} = 77.93$, suggesting that the mixed Poisson regression model fits adequately. Moreover, as discussed in Section 2.7.3, the residual analysis shows that there are a few influential observations and outliers. For example, the 12th observation is an influential observation corresponding to the company which spent \$33.8 million on R&D for 59 patent applications. On omitting the 12th observation, the new parameter estimates become

$$\begin{aligned}
\hat{\alpha}_1 &= (13.7407, 7.7893, -0.8036), \\
\hat{\alpha}_2 &= (0.4344, 1.8847, -0.2071), \\
\hat{\alpha}_3 &= (0.7056, 0.5177, 0.0744), \\
\hat{p}_1 &= 0.1653, \quad \hat{p}_2 = 0.1929, \quad \text{and} \quad \hat{p}_3 = 0.6418.
\end{aligned}$$

Note that the changes in the parameter estimates of the first component are relatively large, while the changes in the other parameter estimates are not significant.

The fitted mixed Poisson regression model suggests that patent counts are generated by three underlying Poisson distributions with rates defined by three different R&D productivity functions, respectively,

$$\lambda_1(x_i^{(r)}, \alpha_1) = \exp[-16.223 + 9.309 \log(R\&D_i) - 1.014\ (\log(R\&D_i))^2],$$

$$\lambda_2(x_i^{(r)}, \alpha_2) = \exp[0.590 + 1.780 \log(R\&D_i) - 0.196 (\log(R\&D_i))^2]$$

$$\text{and } \lambda_3(x_i^{(r)}, \alpha_3) = \exp[0.703 + 0.518 \log(R\&D_i) + 0.076 (\log(R\&D_i))^2].$$

Note that since the above three rate functions are conditional on the three underlying states respectively, the coefficients in these functions should be interpreted as the effects on conditional mean. For instance, $\hat{\alpha}_{12} = 9.309$ is the $\log(R\&D)$ effect on patents when a firm is in state one.

The three dotted lines in Figure 2.6 represent the curves of the above functions respectively. The implied R&D elasticities (derivatives with respect to $\log(R\&D)$) are $9.309 - 2.028 \log(R\&D)$, $1.780 - 0.392 \log(R\&D)$ and $0.518 + 0.152 \log(R\&D)$, suggesting that returns to scale differ across components.

Note that when we fit the data by the usual Poisson regression, which fails to account for the excess variation, the quadratic term is not significant. (The difference in the log likelihood between the usual Poisson regression models with and without the quadratic term is 0.45 and $\chi^2_{1,0.99} = 6.634 > 2 * 0.45 = 0.9$.) If this were the correct model, we would conclude that economies of scale do not vary significantly with the size of the firm's R&D program. The mixture model estimated above indicates, however, that the quadratic term *is* significant in terms of likelihood ratio test. (The difference in the log likelihood between 3-component mixture models with and without the quadratic term is 6.56 and $\chi^2_{3,0.99} = 11.345 < 2 * 6.56 = 13.12$.) This result exemplifies that overdispersion in the usual Poisson regression may result in too large standard error estimates, and subsequently reject too many items in the usual Poisson regression.

If we postulate three different states in terms of the above three different forms of the Poisson rates, a firm has 0.1819 probability of being in state 1, 0.1773 of being in state

2 and 0.6408 of being in state 3. Based on the estimated posterior probabilities defined in (2.20), we identify each firm with one of the three states. Figure 2.6 displays this classification in which a firm is identified with a state if the estimated posterior probability of the firm's being in that state has the largest value. The maximum estimated posterior probabilities always exceeds 0.5 in this application. Note that those observations marked as "1" form a group characterized by $\lambda_1(x_i^{(r)}, \alpha_1)$, those marked as "2" by $\lambda_2(x_i^{(r)}, \alpha_2)$, and those marked as "3" by $\lambda_3(x_i^{(r)}, \alpha_3)$.

For the purpose of comparison, we fit the data to three widely used quasi-likelihood models. The first assumes a variance function $Var(Y_i) = \sigma^2 E(Y_i)$, and the second $Var(Y_i) = E(Y_i) + \sigma^2 E(Y_i)^2$. Note that the negative binomial model has such a mean-variance relationship. Further, the parameter estimates under the negative binomial model may not be significantly different from those obtained by the quasi-likelihood, though the former may be more efficient (Lawless, 1987). The third assumes that $E(Y_i) = \lambda_i$ is a random variable, and that $\log(\lambda_i) = x_i\beta + \epsilon_i$ where x_i are covariates, β are unknown regression parameters, and ϵ_i are random error terms having mean 0 and a constant unknown variance σ^2. The unknown parameter σ^2 in these models is called unexplained variance. Estimation for these models is discussed by McCullagh and Nelder (1989) and Breslow (1984).

The results of parameter estimates and standard errors are given in Table 2.10. Computing the t-statistic (estimated coefficient/standard error) and comparing the mixed Poisson regression model with the quasi-likelihood, we find that all three quasi-likelihood models may underestimate the effects of R&D innovation. For example, the absolute values of the t-statistics of the estimated coefficient for $(\log(R\&D))^2$ are 0.398, 3.418 and 1.554 for the quasi-likelihood model I, II and III respectively, while the values of the same coefficient in the mixed Poisson regression model are 4.955, 4.560 and 4.314 for the first, second and third components.

In summary, we have applied the mixed Poisson regression model to analyze the relationship between technological innovation and R&D research at firm level. The patent data are well fitted by the 3-component mixed Poisson regression model with constant mixing probabilities and Poisson rates defined by quadratic functions in $\log(R\&D)$. This shows that both covariates $\log(R\&D)$ and $(\log(R\&D))^2$ are significant predictors of the number of patent applications. On the other hand, the covariate $(\log(R\&D))^2$ is not significant in the usual Poisson regression model which may not be justifiable here because of overdispersion. The goodness-of-fit test shows that there is no significant evidence of lack of fit in the mixed Poisson regression model. In addition, the residual analysis identifies outliers and influential observations in terms of the fitted model. According to the fitted model, the firms are classified into three categories, each characterized by a Poisson rate function. Note that the significance of the parameter estimates of the mixed Poisson regression model is quite different from that obtained by the quasi-likelihood methods for dealing with extra-Poisson variation.

2.8.2 Seizure Frequency in a Clinical Trial

The timing and circumstances of epileptic seizure recurrence are a source of apprehension for the patients and a mystery for the neurologists. Thus there have been many clinical studies of different treatments for reducing occurrence of epileptic seizures, and accordingly various methods used to assess a reduction in seizure frequency (e.g., Wilensky, et al., 1981, Hopkins, et al., 1985, Milton, et al., 1987, Gram, 1988, and Albert, 1991). Some of these methods like the percentage of patients "improved," "unchanged," or "worse" are rather subjective. This kind of the methods cannot be used to form anything other than an impressionistic opinion of the value of a treatment unless formal criteria for evaluating the significance of changes in the various parameters are first defined; others are designed for particular situations. For instance, Hopkins et al. (1985) first proposed

a two-state Markov mixture model to describe apparent clustering among daily seizure counts for epileptics. They assumes that at each state the number of seizures is generated by a Poisson distribution, and that transitions between the two states are governed by a Markov chain. Albert (1991) and Le, Leroux and Puterman (1993) presented two different algorithms to find the estimates of the parameters in the model. All these methods do not directly include treatment effects as covariates in model building so that the treatment effects may be difficult to assess.

In this subsection we analyze data from a clinical trial carried out at British Columbia's Children's Hospital which investigated the effect of intravenous gammaglobulin (IVIG) on suppression of epileptic seizures. Subjects were randomized into two groups. After a four week (28 days) baseline observation, the treatment group received monthly infusion of IVIG while the control group received "best available therapy". The primary end point of the trial was daily seizure frequency. The principal data source was a daily seizure diary which contained the number of hours of parental observation and the number of seizures of each type during the observation period.

We use Poisson regression to analyze a series of myoclonic seizure counts from a single subject receiving IVIG. Data extracted from the seizure diary was the daily counts, y_i, and the hours of parental observation t_i for the ith day. Figure 2.7 gives the time plot of daily seizure counts. As covariates we use treatment (x_{i1}), trend (x_{i2}) and treatment-trend interaction (x_{i3}), where

$$x_{i1} = \begin{cases} 1 & \text{if there is a treatment } (\ i > 28) \\ 0 & \text{otherwise, } (i \leq 28\ ) \end{cases} \tag{2.42}$$

$$x_{i2} = \log(i) \tag{2.43}$$

$$\text{and} \quad x_{i3} = x_{i1}x_{i2}. \tag{2.44}$$

The second column in Table 2.11 reports results of fitting the data using the usual

Poisson regression with covariates defined in (2.42), (2.43) and (2.44), and a log link function. The data are overdispersed with respect to the Poisson distribution, since each of the overdispersion tests is highly significant ($P_a = 16.18$, $P_b = 16.22$ and $P_c = 36.33$). This suggests the inadequacy of the usual Poisson regression model.

We apply the mixture model assuming that

(1) each daily observed seizure count, y_i, is associated with time exposure (observation hours), t_i, and covariates $x_i^{(m)} = (1)$ and $x_i^{(r)} = (x_{i1}, x_{i2}, x_{i3})$, where x_{i1}, x_{i2} and x_{i3} are defined in (2.42), (2.43) and (2.44). Note that we assume constant mixing probabilities here because it is believed that the likelihood of being a particular state is a constant for a patient;

(2) daily seizure counts are independent and follow a mixed Poisson regression model with means equal to the product of observation time (t_i) and the Poisson rate (number of seizures per hour). Rates are specified by exponential link functions

$$\lambda_j(x_i^{(r)}, \alpha_j) = \exp(\alpha_{j0} + \alpha_{j1}x_{i1} + \alpha_{j2}x_{i2} + \alpha_{j3}x_{i3}).$$

where $i = 1, \ldots, 140$, $j = 1, \ldots, c$, and c is the number of components in the mixture model. This model allows the treatment, trend and interaction of the treatment and trend to affect the Poisson rate, and the regression coefficients to vary across components.

Table 2.12 provides the results of fitting these models. Among the three saturated mixture models, both AIC and BIC suggest a 2-component model. Within the class of two component models, we can carry out likelihood ratio tests for treatment, trend and interaction effect respectively. For example, to test interaction effect, i.e., $H_0 : \alpha_{j3} = 0$ for $j = 1, 2$, we find that the likelihood ratio statistic equals $2 * (426.21 - 376.18) = 100.06 > \chi^2_{2,0.99} = 9.21$. This suggests a highly significant treatment-trend interaction. The model we finally select is the 2-component saturated mixture.

After fitting the 2-component mixed Poisson regression model to the data, the Pearson goodness-of-fit statistic X^2 is 134.0 with 131 degrees of freedom. This value does not exceed the upper 95% critical point of the χ^2-distribution on 131 degrees of freedom, $\chi^2_{131,0.95} = 158.7$, suggesting that the mixed Poisson regression model fits adequately. Furthermore, the Pearson, deviance and likelihood residuals from the fitted model are calculated and displayed in Figure 2.10, Figure 2.11 and Figure 2.12 respectively. Figure 2.10 shows that the Pearson residuals may not be approximately normal. On the other hand, both Figure 2.11 and Figure 2.12 show that the deviance residuals and likelihood residuals are very similar to each other, and that the 61st observation is far distant from the remaining observations in both plots, suggesting that it may be an outlier. On omitting this observation, the deviance reduction is $r^2_{L61} = (-0.314)^2 = 9.86$. This means that the 61st observation has great impact on the overall fit of the mixed Poisson regression model to the data.

For detection of influential observations, the average relative coefficient changes w_i are calculated and displayed in Figure 2.13. Clearly, the 6th observation is the only influential observation suggested by the plot. On omitting the 6th observation, the average relative coefficient change for each parameter estimate is about 20%, and the new parameter estimates become

$$\begin{aligned}
\hat{\alpha}_1 &= (2.2701, 1.8800, -0.2006, -0.6373), \\
\hat{\alpha}_2 &= (2.0045, 7.4989, -0.2444, -2.3026), \\
\hat{p}_1 &= 0.2740 \quad \text{and} \quad \hat{p}_2 = 0.7260.
\end{aligned}$$

Note that the changes in the parameter estimates of the first component is relatively larger than that in the other parameter estimates. After excluding the 6th observation, we reanalyze the data by fitting to the Poisson regression and 2-3 component mixed Poisson regression models, and select the same mixed Poisson regression model with the above

new parameter estimates. In fact, the values of AIC for the Poisson regression and the saturated 2-3 component mixed Poisson regression models are -576.4, -379.7 and -383.8 respectively, and the values of BIC are -582.3, -392.9 and -404.4 respectively. Further, the likelihood ratio tests lead to the choice of the saturated 2-component mixed Poisson regression model. Hence, residual analysis identifies possible outliers and influential observations in terms of the mixed Poisson regression model.

We now interpret the fitted model. In it the mixing probabilities equal 0.2761 and 0.7239 and the respective rates are

$$\lambda_1(x_i^{(r)}, \alpha_1) = \exp[2.8450 + 1.3020x_{i1} - 0.4063x_{i2} - 0.4309x_{i3}]$$

$$\text{and } \lambda_2(x_i^{(r)}, \alpha_2) = \exp[2.0704 + 7.4318x_{i1} - 0.2707x_{i2} - 2.2762x_{i3}].$$

Note that since the above two rate functions are conditional on the two underlying states respectively, the coefficients in these functions should be interpreted as the effects on on conditional mean. For example, $\hat{\alpha}_{12} = 1.3020$ is the treatment effect when the patient is in state one, while $\hat{\alpha}_{22} = 7.4318$ is the treatment effect in state two.

Figure 2.8 provides the estimated hourly seizure rate corresponding to each component (the solid line is the rate for component one and the dotted line for component two) and the observed hourly seizure rate y_i/t_i. Observe that with the treatment both the hourly rates are lower and the trend is less steep than at baseline, suggesting that this patient benefited from IVIG therapy. Figure 2.9 depicts the estimated mean $E(Y_i)$ (the solid line) and variance $Var(Y_i)$ (the dotted line) for the fitted model obtained through (2.11) and (2.12). Observe that with the treatment the variance becomes much closer to the mean, suggesting the patient's situation becomes more stable. Further, the variance exceeds the mean throughout, with the greatest difference in the baseline period. The "bumpiness" in these quantities is due to the non-constant exposure. Note also that there is no obvious

parametric relationship between the estimated mean and variance.

We note that the clinical investigators conducting this study found the two component model plausible. They said that they have observed subjects to have "bad days" and "good days" with no obvious explanation of this effect. We believe our model captures this aspect of the data and by doing so provides a clinically meaningful explanation of overdispersion. Note that Figure 2.8 also classifies the days in terms of the estimated posterior probabilities. Those observations marked as "1" form a group which is characterized by the Poisson rate function $\lambda_1(x_i^{(r)}, \alpha_1)$, while those marked as "2" form another group which is characterized by $\lambda_2(x_i^{(r)}, \alpha_2)$. We may regard $\lambda_1(x_i^{(r)}, \alpha_1)$ as the Poisson regression specification for group one, and $\lambda_2(x_i^{(r)}, \alpha_2)$ for group two. In this sense, our model consists of two Poisson regression models, each describing the seizure frequency rate on "bad days" and "good days" respectively.

For the purpose of comparison, we also fit the data to the three quasi-likelihood models defined in Section 2.8.1. Table 2.11 reports parameter estimates for these methods. From Table 2.11 we find that using different methods for overdispersion may lead to either different parameter estimates or different standard errors or both. For instance, the coefficient estimate for treatment effect is 4.132 by model I, 4.656 by model II, 3.757 by model III, and 1.3020 for component 1 and 7.4132 for component 2 by our mixture. Further, the ratio of estimate to standard error for trend is 5.8535 under Method I, 4.2559 under Method II, 4.5145 under Method III, and 2.6550 for component 1 and 14.587 for component 2 under our mixture. This implies that these methods disagree to the significance of background trend effect. Compared with the three methods for overdispersion, our mixture model has smaller confidence intervals for parameter estimates.

In this example, we have analyzed the series of myoclonic seizure counts from a clinical trial. The data are well fitted by 2-component mixed Poisson regression model

with constant mixing probabilities and Poisson rates depending on covariates treatment, trend and treatment-trend interaction. The goodness-of-fit test suggests that there is no significant evidence of lack of fit in the model. In addition, the residual analysis identifies influential observations and outliers. According to this model, the patient may have two states of seizure frequency rate, which describe "bad days" and "good days" situations respectively. Comparing with the quasi-likelihood methods, the mixed Poisson regression model gives smaller confidence intervals of parameter estimates. Note that both parameter and standard error estimates under the mixed Poisson regression model differ from those obtained by the quasi-likelihood method.

2.8.3 Terrorist Bombing

We analyze data consisting of a time series of the number of international terrorist bombing episodes (Roberts, 1991, p.432). Roberts (1991) notes that the data do not behave as a single homogeneous series, and suggests that an indicator variable be used to model a level shift for the last seven years. This is reinforced by the time plot (Figure 2.14) which suggests that there might have been a change in rate in 1973.

We first apply the usual Poisson regression with an intercept, trend variable $\log(i)$ and a step variable s_i defined by

$$s_i = \begin{cases} 0 & \text{for } i < 60 \\ 1 & \text{otherwise.} \end{cases} \tag{2.45}$$

Note that defining the step variable as above, we assume that the step change happened in the 60th month. The trend variable is insignificant, and regression estimates are 0.7498(0.0887) for intercept and 1.158(0.0981) for the coefficient of the step variable. The deviance for the model is 368.1 with 142 degrees of freedom. Note that the data are overdispersed in terms of the Poisson regression, since all three score tests for overdispersion (Dean, 1992) are highly significant ($P_a = P_b = 13.84$, and $P_c = 14.59$).

We apply a mixed Poisson model in which

(1) the monthly terrorist bombing count, y_i, is associated with exposure t_i and covariates $x_i^{(r)} = (1)$ and $x_i^{(m)} = (1, \log(i), s_i)$, where $t_i = 1$ (one month) and s_i is defined by (2.45). Note that the covariate $\log(i)$ represents a trend, and Poisson rates are constant;

(2) y_i, $i = 1, \ldots, 144$, are independent and follow a mixed Poisson model with rates, λ_j, and mixing probabilities defined as

$$p_j(x_i^{(m)}, \beta) = \frac{\exp[\beta_{j0} + \beta_{j1}\log(i) + \beta_{j2}s_i]}{\sum_{k=1}^{c-1}\exp[\beta_{k0} + \beta_{k1}\log(i) + \beta_{k2}s_i] + 1}, \quad (j = 1, ..., c-1),$$

$$\text{and} \quad p_c(x_i^{(m)}, \beta) = 1 - \sum_{j=1}^{c-1} p_j(x_i^{(m)}, \beta),$$

where $i = 1, \ldots, 144$ and c is the number of components in the mixture.

This model allows mixing probabilities to depend on the trend variable and step change and to vary between different forms of them.

Table 2.13 provides the results of model fitting. Among the four saturated mixture models, both AIC and BIC suggest a 3-component mixture model. To test whether the trend effect is significant, we first compare the mixture with covariates including an intercept and the step change with the 3- component saturated mixture. The difference in log-likelihood between the two is 0.89, and the chi-square test statistic is $2*0.89 = 1.78$ with 2 degrees of freedom. Hence the trend effect is not significant based on the usual likelihood ratio test. Similarly, comparing the mixture without covariate with the one with the step change variable in covariate, we find that the step change is significant based on the likelihood ratio test. (The chi-square test statistic is 61.62 with 2 degrees of freedom.) Further, we can compare two non-nested mixtures with only step change

variable in covariates and only the trend variable respectively using either AIC or BIC. Clearly, the former has bigger AIC and BIC values. According to the model selection procedure, we finally choose, within the class of 3-component mixtures, the model with a step change in the mixing probabilities.

After fitting the 3-component mixed Poisson regression model to the data, the Pearson goodness-of-fit statistic X^2 is 134.7 with 137 degrees of freedom. This value does not exceed the upper 95% critical point of the χ^2-distribution on 131 degrees of freedom, $\chi^2_{137,0.95} = 165.3$, suggesting that there is no evidence of lack of fit. Furthermore, the Pearson, deviance and likelihood residuals from the fitted model are calculated and displayed in Figure 2.16, Figure 2.17 and Figure 2.18 respectively. Figure 2.16 shows that the Pearson residuals may not be approximately normal. On the other hand, Figure 2.17 and Figure 2.18 show that the deviance residuals and likelihood residuals are very similar to each other, and that the 7th observation is far distant from the remaining observations in both plots, suggesting that it may be an outlier. On omitting this observation, the deviance reduction is $r^2_{L7} = (3.121)^2 = 9.741$. This means that the 7th observation has great impact on the overall fit of the mixed Poisson regression model to the data.

For detection of influential observations, the average relative coefficient changes w_i are calculated and displayed in Figure 2.19. Clearly, the 7th observation is the only influential observation suggested by the plot. On omitting the 7th observation, the average relative coefficient change for each parameter estimate is about 586%, and the new parameter estimates become

$$\begin{aligned}
\hat{\beta}_1 &= (30.28, -30.10), \\
\hat{\beta}_2 &= (27.56, -26.05), \\
\hat{\lambda}_1 &= 1.6874, \quad \hat{\lambda}_2 = 6.3577 \quad \text{and} \quad \hat{\lambda}_3 = 14.239.
\end{aligned}$$

Note that the changes in the regression parameter estimates of the mixing probabilities

are very significant. This may be due to the fact that after excluding the 7th observation, the first 60 observations are all generated by the first two components. Hence the mixing probability of the third component is almost zero. In this case, the parameter estimates may lead infinity because they are on the boundary of the parameter space, as it usually happens in logistic regression. Note also that the Poisson rates do not change significantly, suggesting that the 7th observation has great influence on the mixing probabilities rather than on the Poisson rates. The residual analysis confirms that the fitted model is adequate. We interpret the fitted mixed Poisson regression model as follows.

In it the mixing probabilities are

$$p_1(x_i^{(m)}, \beta) = \frac{\exp(4.0231 - 3.8535 s_i)}{\exp(4.0231 - 3.8535 s_i) + \exp(1.3141 + 0.1741 s_i) + 1},$$

$$p_2(x_i^{(m)}, \beta) = \frac{\exp(1.3141 + 0.1741 s_i)}{\exp(4.0231 - 3.8535 s_i) + \exp(1.3141 + 0.1741 s_i) + 1}$$

$$\text{and} \quad p_3(x_i^{(m)}, \beta) = \frac{1}{\exp(4.0231 - 3.8535 s_i) + \exp(1.3141 + 0.1741 s_i) + 1}.$$

and the Poisson rates are

$$\lambda_1 = 1.6864, \quad \lambda_2 = 6.3611 \quad \text{and} \quad \lambda_3 = 14.044.$$

This model suggests that the mixing probabilities have a jump. During the first 60 months, the number of episodes follows one of three Poisson distributions with a low rate of 1.6864 (episodes per month) with probability of 0.9221, a medium rate of 6.3611 with probability 0.0614 and a high rate of 14.044 with probability 0.0164 respectively. After December 1972, the data follow one of the same Poisson distributions, however the probabilities have changed to 0.1791, 0.6697 and 0.1512 respectively. This indicates that terrorist bombing incidents become significantly more frequent between 1973 and 1979. Furthermore, the mixture model suggests that the time trend (monthly index) is not significant, suggesting that rates are stable in these periods.

If we postulate three levels of terrorist bombing corresponding to the three different Poisson rates, each month occupies one of the levels according to the mixing probabilities. Based on the estimated posterior probabilities defined in (2.20), we identify each observation with a level if its estimated posterior probability of being at that level is greater than 0.5. Figure 2.15 classifies months in this way. Note that the high intensity component counts for the large number of episodes in July 1968 as well many past 1973 data parts.

From the fitted model, we find the estimated mean and variance are 2.18 and 6.69, respectively, for the first five years, and 5.82 and 19.5 for the last seven years. Clearly, the mixed Poisson model accounts for overdispersion.

Note that we also fit the data using mixed Poisson regression model with a step change in the rate, and have found that the above model fits better.

In summary, the terrorist bombing data have been fitted by the 2-component mixed Poisson regression model with constant Poisson rates and mixing probabilities depending on a step change. This means that since July 1968 terrorist bombing have become more intensive because of a likelihood of being a higher bombing rate. The goodness-of-fit test shows that there is no significant evidence of lack of fit. In addition, the residual analysis identifies one observation which is not only an outlier but also an influential observation in terms of the fitted model.

2.8.4 Accidents in Worksites

There have been many studies of the relationship between alcohol and accident injuries (e.g., McDermott, 1977; Dietz and Baker, 1974; Hingson and Howland, 1987; and Wechsler et al., 1969). Some of these studies established a link between alcohol and accidental injuries (McDermott, 1977), but others have not. Particularly, there is no strong evidence implicating alcohol in workplace injuries. Some methodological issues associated

with these studies include data collection, alcohol measurement and appropriate statistical models. Webb et al. (1994) conducted a study to analyze the relationship between problem drinking and industrial workplace injuries. They collected data from 470 employees of a large industrial plant manufacturing metal products in the Hunter Valley region of New South Wales, Australia, employed during period May 1985 to July 1986. Problem drinking was measured by the Mortimer-Filkins test, which was devised initially to detect alcohol problem among persons charged with drunk-driving (Mortimer et al., 1971). The range of the test scores (MFts) in the data varies from -3 to 37. The numbers of work injuries were obtained from medical reports completed for all injuries reported to the medical center by study participants, for a period of 12 months from the time of administration of the questionnaire to each study participant. The data also contain socio-demographic measures including age and job satisfaction. A question of interest here is to find significant predictors of work injuries.

A review of studies on the relationship between alcohol and work injuries revealed that the evidence is contradictory and that many of the studies contain methodological flaws (Webb et al., 1994). As a standard method for count data analysis, we use Poisson regression by defining the number of work injuries in subject i as Y_i and including covariates:

$$x_{i1} = \log(age_i) \tag{2.46}$$

$$x_{i2} = \begin{cases} 1 & \text{if individual } i \text{ has low level of job satisfaction} \\ 0 & \text{otherwise,} \end{cases} \tag{2.47}$$

$$x_{i3} = \log(MFts_i + 10) \tag{2.48}$$

$$\text{and} \quad x_{i4} = x_{i1}^2. \tag{2.49}$$

Thus the model for Poisson mean λ_i is

$$\log(\lambda_i) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4}.$$

Note that we add a constant 10 in (2.48) so that $MFts_i + 10 > 0$ and the log-transfer can be applied.

The first row in Table 2.14 reports the results of fitting the data using usual Poisson regression. Comparing the t-statistics (parameter estimate/standard error), all covariates except x_{i4} are highly significant. However these results may be misleading because the data are seriously overdispersed. The overdispersion score test statistic P_a has a value of 24.33 which was compared to the $N(0,1)$ reference value, and suggests inadequacy of the usual Poisson regression model.

To apply the mixed Poisson regression model, we assume that

(1) the number of work injuries for individual i is associated with covariates $x_i = (x_i^{(m)}, x_i^{(r)})$ with $x_i^{(m)} = (1, x_{i1}, x_{i2}, x_{i3})'$, $x_i^{(r)} = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})$, where x_{i1}, x_{i2}, x_{i3} and x_{i4} are defined by (2.46), (2.47), (2.48) and (2.49) respectively. Note that we choose $t_i = 1$ for all i;

(2) injury counts of different individuals are independent and follow a mixed Poisson regression model with rates (number of work injuries per year) given by the link functions

$$\lambda_j(x_i^{(r)}, \alpha_j) = \exp(\alpha_{j0} + \alpha_{j1}x_{i1} + \alpha_{j2}x_{i2} + \alpha_{j3}x_{i3} + \alpha_{j4}x_{i4})$$

where $i = 1, 2, ..., 470$, $j = 1, 2, ..., c$ and c is the number of components in the mixture.

Table 2.14 shows the results of fitting these models. In order to determine the number of components first, we compare the values of BIC and AIC among the three saturated models. Clearly, both BIC and AIC lead to the choice of 2-component mixture models. Within these 2-component mixtures, we carry out inference using likelihood ratio tests. First we test the hypothesis that the effects of covariates x_{i2}, x_{i3} and x_{i4} are insignificant by comparing the one including only x_{i1} in both mixing probabilities and rates with the saturated 2-component model. Since the chi-square test statistic is $2 * (-897.74 + 903.45) = 11.42 < \chi^2_{8,0.95} = 15.51$, we do not reject the hypothesis at 5% significance

level. This implies that both the level of job satisfaction and Mortimer-Filkins test score do not have significant effects on mixing probability and Poisson rates.

Then we test whether the effect of age (x_{i1}) is insignificant in the mixing probabilities. Indeed, age is a significant covariate in mixing probabilities because the chi-square test statistic for the corresponding hypothesis is $2 * (-899.15 + 906.31) = 14.32 > \chi^2_{1,0.95} = 3.84$.

For Poisson rates, the age covariate x_{i1} is also highly significant in the rates because the corresponding test statistic is $2 * (-903.45 + 909.57) = 12.24 > \chi^2_{2,0.95} = 5.99$. Finally we test the hypothesis of a common slope for both components, i.e., $\alpha_{11} = \alpha_{21}$. Indeed this hypothesis is valid at 5% significance level because the test statistic is $2*(-903.45+903.48) = 0.06 < \chi^2(1, 0.95) = 3.84$. Therefore we choose the 2-component mixture model with the age covariate in mixing probabilities and Poisson rates with the common coefficient. This model fits the data best.

After fitting the 2-component mixed Poisson regression model to the data, the Pearson goodness-of-fit statistic X^2 is 510.8 with 465 degrees of freedom. This value does not exceed the upper 95% critical point of the χ^2-distribution on 465 degrees of freedom, $\chi^2_{465,0.95} = 516.27$, suggesting that there is no evidence of lack of fit in the mixed Poisson regression model. Furthermore, the Pearson, deviance and likelihood residuals from the fitted model are calculated and displayed in Figure 2.21, Figure 2.22 and Figure 2.23 respectively. Figure 2.21 shows that the Pearson residuals may not be approximately normal. On the other hand, Figure 2.22 and Figure 2.23 show that the deviance residuals and likelihood residuals are very similar to each other, and that the numbers of the possible outliers in these two plots are the same, with the 72th observation having the largest values of deviance and likelihood residuals. On omitting the 72th observation, the deviance reduction is $r^2_{L72} = (3.488)^2 = 12.166$. This means that this observation has great impact on the overall fit of the mixed Poisson regression model to the data.

For detection of influential observations, the average relative coefficient changes w_i are calculated and displayed in Figure 2.24. Clearly, the plot shows that there are a couple of influential observations with the 434th observation having the largest value (0.417). On omitting the 434th observation, the average relative coefficient change for each parameter estimate is about 42%, and the new parameter estimates become

$$\begin{aligned}
\hat{\alpha}_1 &= (-1.1505, 0.2566), \\
\hat{\alpha}_2 &= (0.5850, 0.2566) \quad \text{and} \\
\hat{\beta}_1 &= (-6.5083, 1.9982).
\end{aligned}$$

Note that the changes in the regression parameter estimates of the Poisson rates, especially the common regression parameter, are relatively larger than that in mixing probabilities. This suggests that the 434th observation has great influence on the Poisson rates rather than on the mixing probabilities. The residual analysis identifies possible outliers and influential observations in terms of mixed Poisson regression model. We now interpret the fitted model as follows.

The chosen mixture model suggests that work injury counts are generated by the two underlying Poisson distributions with rates defined by

$$\lambda_1(x_i^{(r)}, \alpha_1) = \exp(-1.4545 + 0.3431 \log(age_i))$$

$$\text{and} \quad \lambda_2(x_i^{(r)}, \alpha_2) = \exp(0.3066 + 0.3341 \log(age_i)).$$

Also these two distributions are mixed according to the mixing probabilities defined by

$$\begin{aligned}
p_1(x_i^{(m)}, \beta) &= \frac{\exp(-6.8705 + 2.1068 \log(age_i))}{\exp(-6.8705 + 2.1068 \log(age_i)) + 1} \\
\text{and} \quad p_2(x_i^{(m)}, \beta) &= \frac{1}{\exp(-6.8705 + 2.1068 \log(age_i)) + 1}
\end{aligned}$$

According to this model employees may be classified into two groups on the basis of work injury rates. Those in one group have relatively a low baseline risk, and those in group two a high baseline risk. Age, however, has the same effect on both groups. In fact as employees get older, their chances of having a work injury increase. On the other hand, since the mixing probability for group one $p_1(x_i^{(m)}, \beta)$ increases in terms of age, there are more senior employees in the low risk group than young ones. For example, for a 25 year old employee, there is a 47.8% chance of being classified into the low risk group with an accident rate of 0.7 work injuries per year, and 52.2 % chance the high risk group with an accident rate of 4.0 work injuries per year; For a 50 year old employee, there is a 79.8% chance of being classified into the low risk group with an accident rate of 0.9 work injuries per year, and 20.2% chance the high risk group with an accident rate of 5.0 work injuries per year. Figure 2.20 provides the estimated work injury rate corresponding to each group (the solid line is the rate for the low risk group and the dotted line for the high risk group. Note that Figure 2.20 also classifies the employees in terms of the estimated posterior probabilities. Those observations marked as "1" form the low risk group which is characterized by the function $\lambda_1(x_i^{(r)}, \alpha_1)$, while those marked as "2" form the high risk group which is characterized by the function $\lambda_2(x_i^{(r)}, \alpha_2)$.

In this example, we have found that neither the problem drinking measure, the Mortimer-Filkins test score nor the job satisfaction score is a good predictor of workplace injuries. On the other hand, age is a significant predictor of workplace injuries. After taking into account age effects, the accident rates do not depend on Mortimer-Filkins test score and job satisfaction but only on age in the log-linear function. The workplace injury data are well fitted by the 2-component mixed Poisson regression model which consists of two Poisson regression models. According to the model, the employees can be classified into two groups depending on baseline risk and the likelihood of being in one of the baseline groups associated with age. Note also that the inferences differ from

those obtained through the usual Poisson regression analysis. The goodness-of-fit test shows that there is no significant evidence of lack of fit. In addition, the residual analysis identifies several outliers and influential observations in terms of the fitted model.

2.8.5 Aces Salmonella Assay Data

The data in this example were first presented by Margolin et al. (1981) from an Ames salmonella reverse mutagenicity assay, and analyzed by Breslow (1984) and Lawless (1987b) using quasi-likelihood and negative binomial approaches respectively. Table 2.15 shows the number of revertant colonies (y_i) observed on each of three replicate plates tested at each of six dose level of quinoline (d_i).

Lawless (1987b) defined the expected frequency of revertants as

$$E(Y_i \mid d_i) = \lambda(d_i) \equiv \exp(\alpha_0 + \alpha_1 d_i + \alpha_2 \log(d_i + 10)),$$

while Breslow (1984) assumed $E(Y_i \mid d_i) \simeq \lambda(d_i)$. At issue is whether a mutagenic effect is present. This corresponds to testing the hypothesis that $\alpha_2 = 0$. The data are overdispersed relative to Poisson regression with rate defined above, since each of the three tests for overdispersion is highly significant ($P_a = 5.628$, $P_b = 5.656$ and $P_c = 5.607$).

To account for overdispersion, Breslow (1984) assumed a variance function $Var(Y_i) \simeq \lambda(d_i) + \sigma^2\lambda(d_i)^2$, and obtained parameter estimates by using weighted least-squares combined with method of moments. Similarly, Lawless (1987b) fitted the data with a negative binomial model in which the variance function is $Var(Y_i) = \lambda(d_i) + \sigma^2\lambda(d_i)^2$, and obtained parameter estimates by maximum likelihood. Parameter estimates (standard errors) are reported in Table 1.8.5.3.

Our analysis of the data using mixed Poisson regression models follows. We assume

(1) the number of observed revertant colonies, y_i, is associated with covariates $x_i = (1, d_i, \log(d_i + 10))$, and $t_i = 1$;

(2) Y_i are independent and follow a mixed Poisson regression model with Poisson rates

$$\lambda_j(x_i, \alpha_j) = \exp(\alpha_{j0} + \alpha_{j1} d_i + \alpha_{j2} \log(d_i + 10)).$$

where $i = 1, \ldots, 18$ and $j = 1, \ldots, c$.

Table 2.16 shows the results of fitting these models. Among the three saturated models, both AIC and BIC lead to the choice of 2-component mixtures. To test mutagenic effects, we compare the saturated model to the one without covariate $\log(d_i + 10)$ by a likelihood ratio test. Since the chi-square test statistic equals $2 * (68.81 - 60.90) = 15.82 > \chi^2_{2,0.99} = 9.21$, mutagenic effects are significant. Further, the similar regression coefficient estimates for each component in the saturated model suggest common regression coefficients for both components. This is indeed confirmed by the likelihood ratio test (the chi-square test statistic is $2 * 0.01 = 0.02 < \chi^2_{2,0.99} = 9.21$.) Hence we choose to represent the data by the 2-component mixture with common regression coefficients and different intercepts for each component.

The fitted model may be interpreted as follows. In it mixing probabilities equal 0.8173 and 0.1827 and the respective rates are

$$\begin{aligned} \lambda_1(x_i, \alpha_1) &= \exp(1.9094 - 0.00126 d_i + 0.3640 \log(d_i + 10)) \text{ and} \\ \lambda_2(x_i, \alpha_2) &= \exp(2.4768 - 0.00126 d_i + 0.3640 \log(d_i + 10). \end{aligned}$$

This model indicates that mutagenic effects are the same for both components. This model may also be regarded as a Poisson regression with a random intercept following a discrete mixing distribution with 2-points of support. Figure 2.25 shows the classification for the data in which each observation is identified with either of the two components in the mixture according to the estimated posterior probabilities defined by (2.20). This plot may provide a way to visualize overdispersion for the data. From it we conjecture

that the three observations classified with component 2 may be outliers in terms of the Poisson regression model, and that overdispersion may be due to these three observations. In fact, the residual analysis below adds strength to this conjecture.

After fitting the 2-component mixed Poisson regression model to the data, the Pearson goodness-of-fit statistic X^2 is 16.2 with 13 degrees of freedom. This value does not exceed the upper 95% critical point of the χ^2-distribution on 13 degrees of freedom, $\chi^2_{0.95}(13) = 22.36$, suggesting that there is no evidence of lack of fit. Moreover, the Pearson, deviance and likelihood residuals are displayed in Figure 2.26, Figure 2.27 and Figure 2.28 respectively. Figure 2.27 and Figure 2.28 show that the deviance and likelihood residuals are very similar to each other. On the other hand, Figure 2.26 indicates that the Pearson residuals may not be approximately normal.

For detection of influential observations, the average relative coefficient changes w_i are calculated and displayed in Figure 2.29. Clearly, the plot shows that the 12th observation is influential. On omitting the 12th observation, the new estimates of the intercepts in two components are 2.2242 and 2.5460 respectively; the new estimates of the other common regression parameters are -0.00067 and 0.2430 respectively; and the new estimates of the mixing probabilities for the two components are 0.5644 and 0.4356 respectively. Note that the new intercept estimates are very close, suggesting that the data excluding the 12th observation may not be overdispersed. In fact, we fit the data to the Poisson regression model, and find that there is no strong evidence of overdispersion because each of the three overdispersion score test statistics is not significant ($P_a = 1.6142$, $P_b = 1.6132$ and $P_c = 1.8543$). If we use the correction forms of these score test statistics for small samples, $P'_a = 2.1339$, $P'_b = 2.1328$ and $P'_c = 2.3688$. These values are marginal to the normal critical values at critical level $\alpha = 0.5$, suggesting again that there are no strong evidence of overdispersion. Assuming that the data excluding the 12th observation is overdispersed, we also fit the data to the 2 and 3 component mixed Poisson regression

models, and select the (one-component) Poisson regression model because it yields the largest values of AIC and BIC among the three saturated models. That is, the values of AIC and BIC for the Poisson regression and the 2-3 component saturated mixed Poisson regression models are -61.3, -61.4 and -64.4 respectively, and the values of BIC are -62.7, -64.5 and -68.9 respectively. The analysis shows that extra-Poisson variation may be caused by outliers in terms of Poisson regression, and that the mixed Poisson regression model may tend to model these outliers by extra components. Note also that the changes in the parameter estimates and corresponding standard errors between the two Poisson regression models with and without the 12th observation may not be very significant, suggesting that the 12th observation may be an outlier in terms of the Poisson regression with the complete data.

From Table 2.17, we note that the regression coefficient estimates, $\hat{\alpha}_1$ and $\hat{\alpha}_2$, do not vary drastically across models, but their standard errors do. For instance, the value of $\hat{\alpha}_2/se(\hat{\alpha}_2)$ changes from $0.3640/0.0665 = 5.4737$ under the mixed Poisson regression model to $0.3110/0.09901 = 3.1411$ under the quasi-likelihood model. Thus, although all four models agree that mutagenic effects are significant, they disagree agree to the significance of the effects. Note that confidence intervals under the mixed Poisson regression model are much smaller than either the quasi-likelihood or negative binomial model. Hence effects are estimated more precisely. For example, an approximate 95% confidence interval for the coefficient of $\log(dose + 10)$ under the mixed Poisson regression is 0.3640 ± 0.1303, 0.3110 ± 0.1941 under quasi-likelihood, and 0.313 ± 0.1701 under the negative binomial model. This suggests that using different models to account for overdispersion may lead to different conclusions.

In this example, we analyzed the data set from an Ames salmonella reverse mutagenicity assay. The data are well fitted by the 2-component mixed Poisson regression model with constant mixing probabilities and Poisson rates as functions of dose level.

Note that the mutagenic effects are the same for both components, while the intercepts in the Poisson rates vary between the two components. The goodness-of-fit test suggests that there are no evidence of lack of fit in the model. In addition, the residual analysis identifies one influential observation. Excluding this observation, the data are not overdispersed. This example suggests that extra-Poisson variation may be caused by the presence of outliers in terms of Poisson regression, and that the mixed Poisson regression may model these outliers by including extra components. This example also illustrates a difference between our approach and the usual approaches for accounting for overdispersion. Since the variance exceeds the mean, methods which correct for this by increasing the variance may lead to less significant regression coefficient estimates. Our approach has a different effect. By attributing overdispersion to the presence of several components, the mixed Poisson regression model estimates coefficient effects with smaller error.

2.9 Tables and Figures in Chapter 2

Table 2.1 The results of the simulations for the mixed Poisson regression models

The First Model												
Comp. i	Poisson Rates						Mixing Probabilities					
	α_{i0}	$E(\hat{\alpha}_{i0})$	$Var(\hat{\alpha}_{i0})$	α_{i1}	$E(\hat{\alpha}_{i1})$	$Var(\hat{\alpha}_{i1})$	β_{i0}	$E(\hat{\beta}_{i0})$	$Var(\hat{\beta}_{i0})$	β_{i1}	$E(\hat{\beta}_{i1})$	$Var(\hat{\beta}_{i1})$
1	2.8	2.7955	0.0424	-2.9	-2.9033	0.1076	1.1	1.1764	0.0845			
2	2.6	2.6146	0.0183	0.4	0.3903	0.0090	0.6	0.5938	0.1050			
3	3.6	3.5986	0.0095	0.2	0.1983	0.0065						

The Second Model												
Comp. i	Poisson Rates						Mixing Probabilities					
	α_{i0}	$E(\hat{\alpha}_{i0})$	$Var(\hat{\alpha}_{i0})$	α_{i1}	$E(\hat{\alpha}_{i1})$	$Var(\hat{\alpha}_{i1})$	β_{i0}	$E(\hat{\beta}_{i0})$	$Var(\hat{\beta}_{i0})$	β_{i1}	$E(\hat{\beta}_{i1})$	$Var(\hat{\beta}_{i1})$
1	0.4	0.3699	0.0268				2.0	1.9238	0.6256	-2.0	-2.0034	0.5776
2	3.0	3.0019	0.0011				-1.4	-1.5175	0.8428	1.5	1.5686	0.4572
3	2.0	1.9882	0.0104									

The Third Model												
Comp. i	Poisson Rates						Mixing Probabilities					
	α_{i0}	$E(\hat{\alpha}_{i0})$	$Var(\hat{\alpha}_{i0})$	α_{i1}	$E(\hat{\alpha}_{i1})$	$Var(\hat{\alpha}_{i1})$	β_{i0}	$E(\hat{\beta}_{i0})$	$Var(\hat{\beta}_{i0})$	β_{i1}	$E(\hat{\beta}_{i1})$	$Var(\hat{\beta}_{i1})$
1	2.8	2.8141	0.0376	-2.9	-2.9461	0.1523	2.0	2.2648	0.9838	-2.0	-2.2528	0.9904
2	3.6	3.5811	0.0069	0.2	0.2119	0.0027	-1.4	-1.4346	0.7530	1.5	1.5776	0.4372
3	2.6	2.5870	0.0187	0.4	0.4062	0.0104						

Table 2.2: The result of a Monte Carlo study on the 2-component mixed Poission regression model with constant mixing probabilities and variable rates --I.

parameter	true value	mean	standard deviation	upper extreme	upper quartile	median	lower quartile	lower extreme
α_1	1.0	0.9885	0.0630	1.0501	1.0000	1.0000	0.9558	0.8972
α_2	-0.5	-0.4793	0.0699	-0.4388	-0.4762	-0.5000	-0.5023	-0.5442
p_1	0.6	0.6037	0.0699	0.8878	0.6688	0.5996	0.5120	0.2917
α_1	2.0	1.9985	0.0420	2.0570	2.0119	2.0000	1.9781	1.9277
α_2	-0.5	-0.4766	0.0585	-0.4224	-0.4689	-0.5000	-0.5000	-0.5327
p_1	0.6	0.6239	0.1437	0.9239	0.7015	0.6055	0.5524	0.3628
α_1	1.0	0.9734	0.0562	1.0613	1.0000	1.0000	0.9347	0.8467
α_2	-2.5	-2.4629	0.1419	-2.4945	-2.5000	-2.5000	-2.5085	-2.5205
p_1	0.6	0.5878	0.0865	0.8029	0.6466	0.5878	0.5276	0.3652
α_1	2.0	1.9932	0.0474	2.0176	2.0000	1.9999	1.9875	1.9704
α_2	-2.5	-2.4828	0.0801	-2.4997	-2.4999	-2.5000	-2.5000	-2.5000
p_1	0.6	0.5961	0.0705	0.7507	0.6475	0.5981	0.5510	0.411728

Table 2.3: The result of a Monte Carlo study on the 2-component mixed Poission regression model with constant mixing probabilities and variable rates --II.

parameter	true value	mean	standard deviation	upper extreme	upper quartile	median	lower quartile	lower extreme
α_1	1.0	0.9906	0.0678	1.0829	1.0105	1.0000	0.9584	0.8813
α_2	-0.5	-0.4820	0.0484	-0.4305	-0.4707	-0.5000	-0.5000	-0.5304
p_1	0.4	0.3975	0.1715	0.6965	0.4759	0.3999	0.3268	0.1045
α_1	2.0	1.9967	0.0430	2.0511	2.0105	2.0000	1.9820	1.9399
α_2	-0.5	-0.4692	0.0657	-0.4298	-0.4699	-0.4993	-0.5000	-0.5432
p_1	0.4	0.4299	0.1683	0.6654	0.4907	0.4087	0.3407	0.1195
α_1	1.0	0.9813	0.0615	1.0547	1.0000	1.0000	0.9625	0.9064
α_2	-2.5	-2.4817	0.0860	-2.4995	-2.4998	-2.5000	-2.5000	-2.5000
p_1	0.4	0.3818	0.0819	0.5619	0.4296	0.3818	0.3323	0.1991
α_1	2.0	1.9938	0.0371	2.0203	2.0000	2.0000	1.9858	1.9657
α_2	-2.5	-2.4763	0.1045	-2.5000	-2.5000	-2.5000	-2.5000	-2.5000
p_1	0.4	0.3965	0.0751	0.5900	0.4483	0.3967	0.3506	0.2086

		p_1	
α_1	α_2	0.6	0.4
1.0	-0.5	63/200	17/200
	-2.5	200/200	198/200
2.0	-0.5	174/200	119/200
	-2.5	200/200	200/200

Table 2.4: The results of the likelihood ratio tests for the hypothesis of $\alpha_2 = 0$ based on the 2-component mixed Poisson regression model–I.

		p_1=0.6		p_1=0.4	
parameter	true value	mean	standard deviation	mean	standard deviation
α_1	1.0	0.9910	0.0849	0.9882	0.0861
α_2	-0.5	-0.2668	0.1362	-0.1611	0.1306
α_1	2.0	1.9985	0.5669	1.9954	0.0546
α_2	-0.5	-0.2708	0.0889	-0.1693	0.0862
α_1	1.0	0.9976	0.0797	0.9903	0.0761
α_2	-2.5	-0.8055	0.2021	-0.4377	0.1598
α_1	2.0	1.9959	0.0500	1.9931	0.0517
α_2	-2.5	-0.8065	0.1789	-0.4536	0.1425

Table 2.5: The results of fitting mixed Poisson regression model to the data from a Monte Carlo study on the 2-component mixed Poisson regression model with constant mixing probabilities and variable rates.

		p_1	
α_1	α_2	0.6	0.4
1.0	-0.5	99/200	47/200
	-2.5	200/200	177/200
2.0	-0.5	181/200	122/200
	-2.5	200/200	200/200

Table 2.6: The results of the likelihood ratio tests for the hypothesis of $\alpha_2 = 0$ based on the 2-component mixed Poisson regression model–II.

Model Number	# of replicates that AIC leads the choice of the right model	# of replcates that BIC leads the choice of the right model	total number of replicates
1	96	100	100
2	87	100	100
3	91	100	100

Table 2.7: The results of model selection based on AIC and BIC values for the Monte Carlo study.

Table 2.8: Poisson regression and overdispersion test statistics for the patent data.

Covariates			log-likelihood	P_a^*	P_b^*	P_c^*
1	$\log(RND)$	$(\log(RND))^2$				
3.155 (0.2466)			-1780.	374.29	374.29	374.29
0.5392 (0.0906)	0.9279 (0.0231)		-316.69	24.47	24.47	28.77
0.6207 (0.1206)	0.8560 (0.0770)	0.0123 (0.0127)	-316.24	24.45	24.45	28.89

* P_a, P_b and P_c are score test statistics which asymptotically follow the standard normal distribution.

Table 2.9: Mixed Poisson regression model estimates for the patent data.

Component number (i)	Poisson rate			Mixing probability	Log-likelihood	AIC	BIC
	α_{i0}	α_{i1}	α_{i2}				
1-component mixture							
1	3.155			1.0	-1780.	-1781.	-1782.
1	0.5392	0.9279		1.0	-316.69	-318.69	-320.93
1	0.6207	0.8560	0.0123	1.0	-316.24	-319.24	-322.61
2-component mixture							
1	1.2046			0.7017	-509.26	-512.26	-515.63
2	4.2599			0.2983			
1	-0.1552	1.0244		0.7058	-219.64	-224.64	-230.26
2	1.7076	0.7240		0.2942			
1	0.3996	0.5567	0.0804	0.7105	-211.61	-218.61	-226.48
2	1.0009	1.2135	-0.0743	0.2895			
3-component mixture							
1	0.9352			0.6177	-356.04	-361.04	-366.66
2	3.3744			0.2215			
3	4.5835			0.1608			
1	-2.3288	1.5232		0.2400	-203.53	-211.53	-226.48
2	1.0771	0.6688		0.1832			
3	0.5829	0.8656		0.5768			
1	-16.233 (3.2566)	9.3086 (1.6424)	-1.0137 (0.2046)	0.1819	-196.97	-207.97	-220.34
2	0.5900 (0.4147)	1.7801 (0.2748)	-0.1961 (0.0430)	0.1773			
3	0.7025 (0.1422)	0.5182 (0.0916)	0.0755 (0.0175)	0.6408			
4-component mixture							
1	0.8643			0.6397	-290.77	-297.77	-305.64
2	2.9917			0.1317			
3	4.0142			0.1288			
4	4.8046			0.0998			
1	-2.2872	1.5144		0.2312	-203.47	-214.47	-226.84
2	0.8841	0.1816		0.0330			
3	0.5898	0.8644		0.5580			
4	1.9882	0.6661		0.1778			
1	0.7759	0.5918	0.0886	0.1715	-194.04	-209.04	-225.90
2	0.7013	0.5423	0.0671	0.4665			
3	-3.6767	2.1199	-0.0659	0.1502			
4	0.3687	1.4506	-0.1180	0.2118			

Table 2.10: Parameter estimates for five models for patent data.

Parameters Estimated	Poisson Regression	Quasi-likelihood I	Quasi-likelihood II	Quasi-likelihood III	Mixed Poisson Regression		
					comp 1	comp 2	comp 3
Intercept	0.6207 (0.1206)	0.6207 (0.2935)	0.9734 (0.1256)	0.6626 (0.1540)	-16.233 (3.2566)	0.5900 (0.4149)	0.7025 (0.1422)
log(R&D)	0.8560 (0.0770)	0.8560 (0.1874)	0.5268 (0.0806)	0.7321 (0.1087)	9.3086 (1.6424)	1.7801 (0.2748)	0.5182 (0.0916)
$(\log(\text{R\&D}))^2$	0.0123 (0.0127)	0.0123 (0.0309)	0.0687 (0.0201)	0.0415 (0.0267)	-1.0137 (0.2046)	-0.1961 (0.0430)	0.0755 (0.0175)
Mixing probabilities					0.1819	0.1773	0.6408
dispersion parameter	1.0	5.917	0.2094	0.4070			

Table 2.11: Parameter estimates for five methods for seizure data.

Parameters Estimated	Poisson Regression	Method I	Method II	Method III	Mixed Poisson Regression	
					Comp 1	Comp 2
Intercept	2.118 (0.0815)	2.118 (0.1897)	2.148 (0.5539)	2.129 (0.3846)	2.8450 (0.2360)	2.0704 (0.0890)
x_1	4.132 (0.3032)	4.132 (0.7059)	4.656 (1.094)	3.757 (0.8322)	1.3020 (0.4904)	7.4318 (0.5095)
x_2	-0.2257 (0.0329)	-0.2257 (0.0766)	-0.2412 (0.2191)	-0.2408 (0.1523)	-0.4063 (0.0909)	-0.2707 (0.0377)
x_3	-1.320 (0.0800)	-1.320 (0.1863)	-1.440 (0.3098)	-1.221 (0.2316)	-0.4309 (0.1385)	-2.2762 (0.1377)
Mixing Probabilities	NA	NA	NA	NA	0.2762	0.7238
Unexplained Variance	1.0	5.4206	0.8631	0.4051	NA	NA

Table 2.12: Mixed Poisson regression model estimates for seizure data.

Component number (j)	Mixing probability	Poisson rate				log likelihood	AIC	BIC
	p_j	α_{j0}	α_{j1}	α_{j2}	α_{j3}			
1-component mixture								
1		2.118	4.132	-0.2257	-1.320	-583.16	-587.16	-593.04
2-component mixture								
1	0.4128	1.2183				-700.10	-703.10	-707.41
2	0.5872	-1.1571						
1	0.3715	1.8959	-1.2761			-462.79	-467.79	-475.14
2	0.6285	1.3777	-3.1018					
1	0.3736	2.9919	-0.4732	-0.4718		-426.21	-433.21	-443.51
2	0.6264	2.1791	-2.3248	-0.3379				
1	0.2761	2.8450 (0.2360)	1.3020 (0.4924)	-0.4063 (0.0909)	-0.4309 (0.1385)	-376.18	-385.18	-398.41
2	0.7239	2.0704 (0.0890)	7.4318 (0.5095)	-0.2707 (0.0377)	-2.2762 (0.1377)			
3-component mixture								
1	0.2742	2.8440	1.2938	-0.4054	-0.4294	-375.29	-389.29	-409.88
2	0.0277	2.0809	-28.767	-0.3928	5.4488			
3	0.6981	2.0694	7.3197	-0.2648	-2.2478			

Table 2.13: Mixed Poisson regression model estimates for terrorist bombing data.

Component number (j)	Mixing probability			Poisson rate	log-likelihood	AIC	BIC
	β_{j0}	β_{j1}	β_{j2}	λ_j			
1-component mixture							
1	NA	NA	NA	4.8125	-480.62	-481.62	-483.10
2-component mixture							
1	4.6461	-0.5822	-2.9447	1.8437	-358.08	-363.08	-370.50
2				8.4545			
3-component mixture							
1	1.7183			1.7224	-378.23	-383.23	-390.65
2	1.5016			6.5450			
3				14.263			
1	10.6997	-2.2057		1.6588	-351.35	-358.35	-368.74
2	-1.7331	0.7280		6.4909			
3				14.331			
1	**4.0231 (1.1500)**		**-3.8535 (0.1712)**	**1.6864 (0.1746)**	-347.42	-354.42	-364.81
2	**1.3141 (1.4417)**		**0.1741 (1.5075)**	**6.3611 (0.5582)**			
3				**14.044 (1.5365)**			
1	3.7129	0.0857	-3.9151	1.6721	-346.53	-355.53	-368.89
2	-2.6160	1.1736	-1.3076	6.3726			
3				14.044			
4-component mixture							
1	1.8888	0.6683	-5.5292	1.5557	-345.53	-358.53	-377.83
2	5.1822	-2.2236	4.6102	2.4601			
3	-2.5787	1.2531	-1.7773	6.4656			
4				14.009			

Table 2.14: Poisson regression and mixed Poisson regression model estimates for the workplace injury data.

g	Mixing probability				Poisson rate					log likelihood	AIC	BIC
j	β_{j0}	β_{j1}	β_{j2}	β_{j3}	α_{j0}	α_{j1}	α_{j2}	α_{j3}	α_{j4}			
Poisson Regression Models												
1	NA	NA	NA	NA	-6.104 (3.085)	4.030 (1.760)	-0.6326 (0.2517)	0.1905 (0.0702)	0.1772 (0.1029)	-1029.7	-1034.7	-1045.1
1	NA	NA	NA	NA	2.347	-0.4530				-1038.7	-1040.7	-1044.8
2-component mixture												
1	-6.0369	2.5200	-0.2908	-0.7256	-39.35	20.83	-2.7608	0.1879	0.02829	-897.74	-911.74	-940.81
2					-2.1322	1.7964	-0.1915	0.0492	-0.0934			
1	-8.8954	2.7054	-0.2539		-42.25	22.07	-2.9241	0.2005	0.2153	-898.79	-911.79	-938.78
2					-2.3575	1.6855	-0.16823	0.0640	0.01803			
1	-8.9136	2.68402			-41.01	21.43	-2.8402	0.2648	0.2005	-899.15	-911.15	-936.07
2					-2.2685	1.6376	-0.1647	0.0990	0.0188			
1	0.9476				-1.3561	0.7861	-0.1942	0.1072	0.3348	-906.31	-917.31	-940.15
2					-4.8160	3.8248	-0.5703	0.1249	0.0126			
1	-8.5642	2.5765			-36.14	19.23	-2.5558			-900.69	-908.69	-925.30
2					-3.1215	2.2312	-0.2570					
1	-6.6585	2.0552			-1.1022	0.2528				-903.45	-909.45	-921.91
2					0.3087	0.3446						
1	-6.8705 (1.6370)	2.1068 (0.4561)			-1.4545 (0.5711)	0.3431 (0.1523)				-903.48	-908.48	-918.86
2					0.3066 (0.5332)							
1	-5.1531	1.4750			-0.5286					-911.21	-915.21	-923.52
2					1.3396							
1	0.8059				-0.1187					-914.78	-917.78	-924.01
2					1.5409							
3-component mixture												
1	0.3376	1.7234	0.9717	-1.9759	-48.39	25.94	-3.3752	-0.0952	-0.8001	-886.10	-909.10	-956.86
2	0.9063	1.3606	0.9713	-1.5373	-10.84	7.0117	-1.0299	0.4992	-0.1443			
3					-3.3993	3.5145	-0.4712	0.2957	-0.4356			

Table 2.15: Number of revertant colonies of salmonella (y_i)

	Dose of quinoline d_i ($\mu g/plate$)					
	0	10	33	100	333	1000
Observed # of colonies	15	16	16	27	33	20
	21	18	26	41	38	27
	29	21	33	60	41	42

Component number (j)	Mixing probability	Poisson rate			log likelihood	AIC	BIC
	p_j	α_{j0}	α_{j1}	α_{j2}			
1-component mixture							
1		2.173	-0.001013	0.3198	-68.13	-71.13	-72.47
2-component mixture							
1	0.6145	3.0779			-68.93	-71.93	-73.27
2	0.3855	3.7112					
1	0.5617	2.9886	0.000188		-68.81	-73.81	-76.04
2	0.4383	3.6428	0.000082				
1	0.8132	1.9125	-0.001247	0.3623	-60.90	-67.90	-71.02
2	0.1868	2.4064	-0.001294	0.3790			
1	0.8173	1.9094	-0.001260	0.3640	-60.91	-65.91	-68.14
2	0.1827	2.4768					
3-component mixture							
1	0.5918	1.8484	-0.001190	0.3640	-60.78	-71.78	-76.68
2	0.3241	2.8535	-0.000154	0.1476			
3	0.0841	5.9320	-0.000100	-0.3895			

Table 2.16: Mixed Poisson regression model estimates for Ames salmonella assay data.

Table 2.17: Parameter estimates for five estimation methods for assay data

Parameters Estimated	Poisson regression		Quasi-Likelihood[1]	Negative Binomial[2]	Mixed Poisson Regression II	
	complete	incomplete[3]			Comp 1	Comp 2
Intercept	2.173 (0.2183)	2.308 (0.2266)	2.203 (0.3634)	2.203 (0.359)	1.9094 (0.2674)	2.3768 (0.2753)
Dose	-0.001013 (0.000245)	-0.000750 (0.000265)	-0.000974 (0.000437)	-0.000980 (0.000381)	-0.001260 (0.000275)	
log(*Dose* + 10)	0.3198 (0.05698)	0.2632 (0.0607)	0.3110 (0.09901)	0.313 (0.0868)	0.3640 (0.0665)	
Mixing Probabilities					0.8173	0.1827
Unexplained Variance	1.0	1.0	0.07181	0.0488		

[1] as appeared in Breslow (1984)

[2] as appeared in Lawless (1987b).

[3] The data excluding the 12th observation.

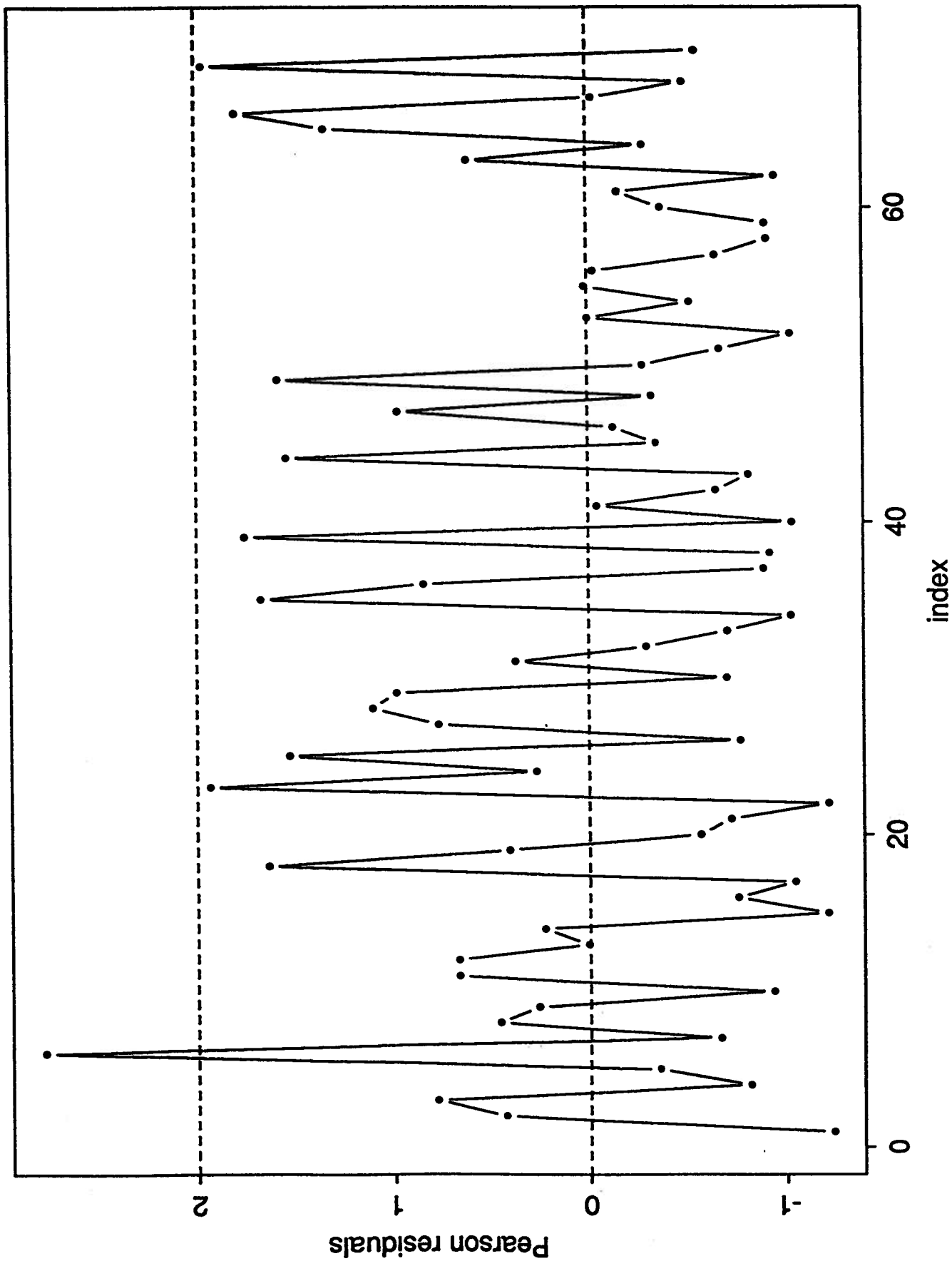

Figure 2.1: The index plot of the Pearson residuals from the fitted 3-component mixed Poisson regression model for the patent data

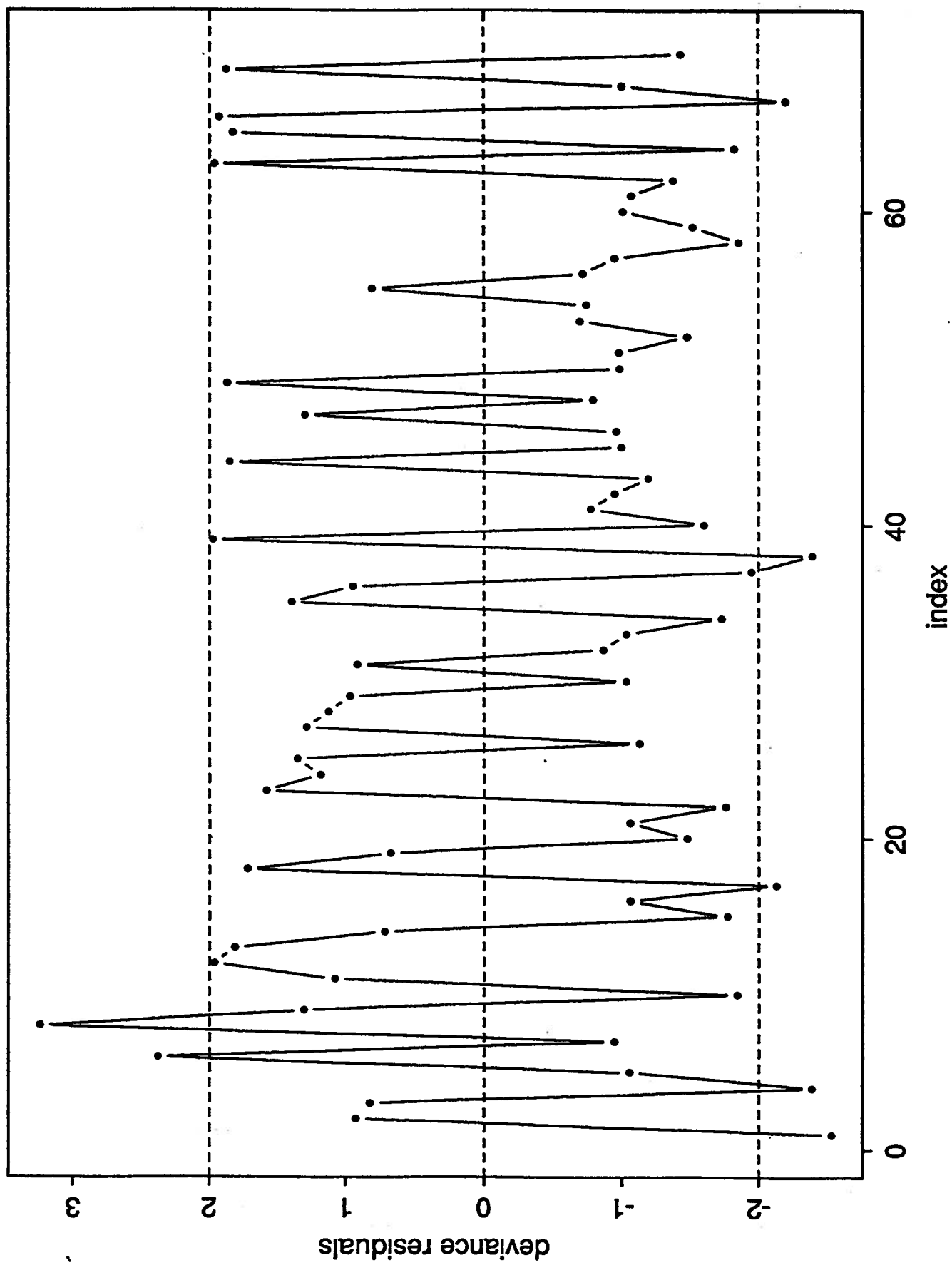

Figure 2.2: The index plot of the deviance residuals from the fitted 3-component mixed Poisson regression model for the patent data.

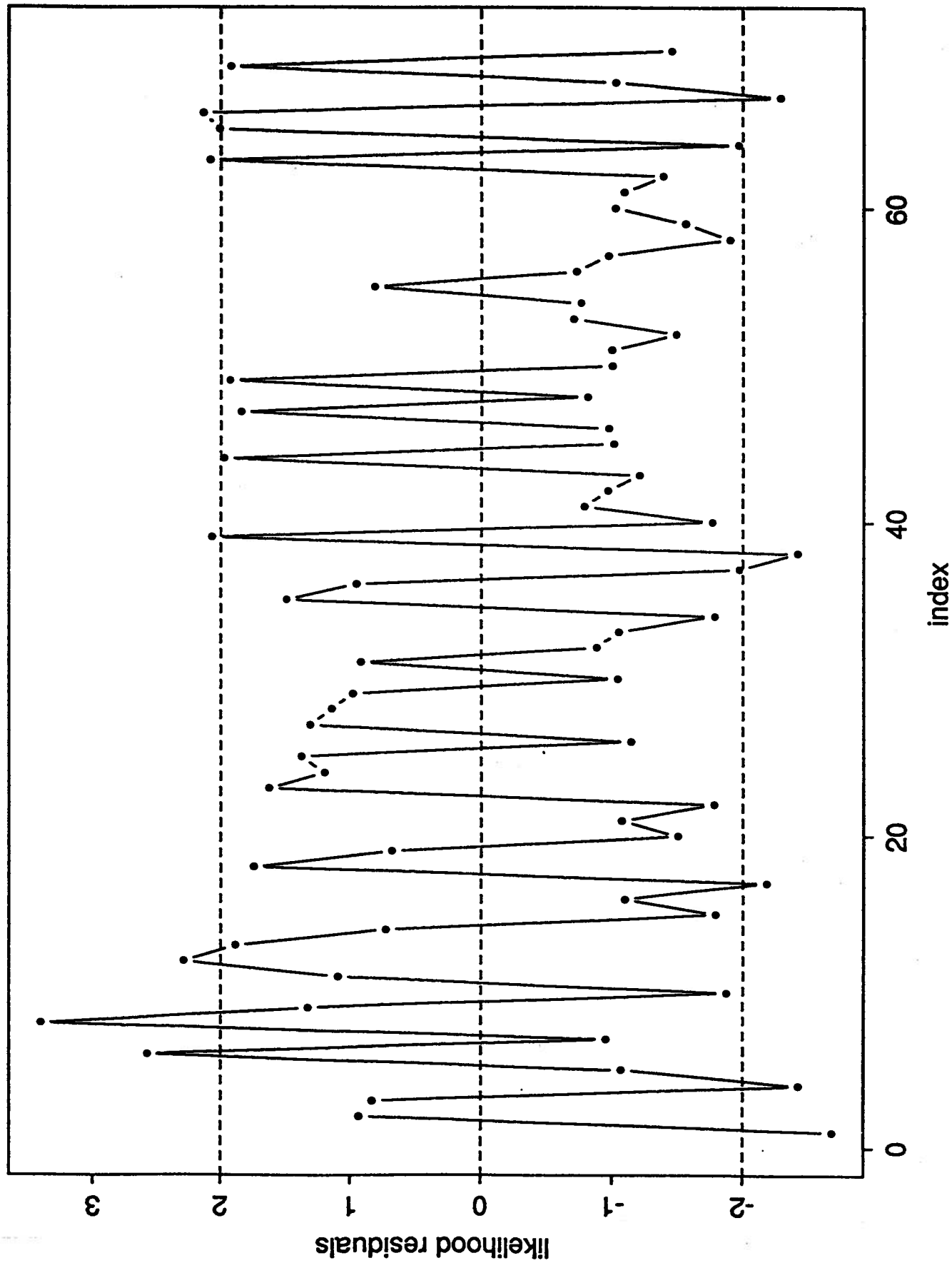

Figure 2.3: The index plot of the likelihood residuals from the fitted 3-component mixed Poisson regression model for the patent data.

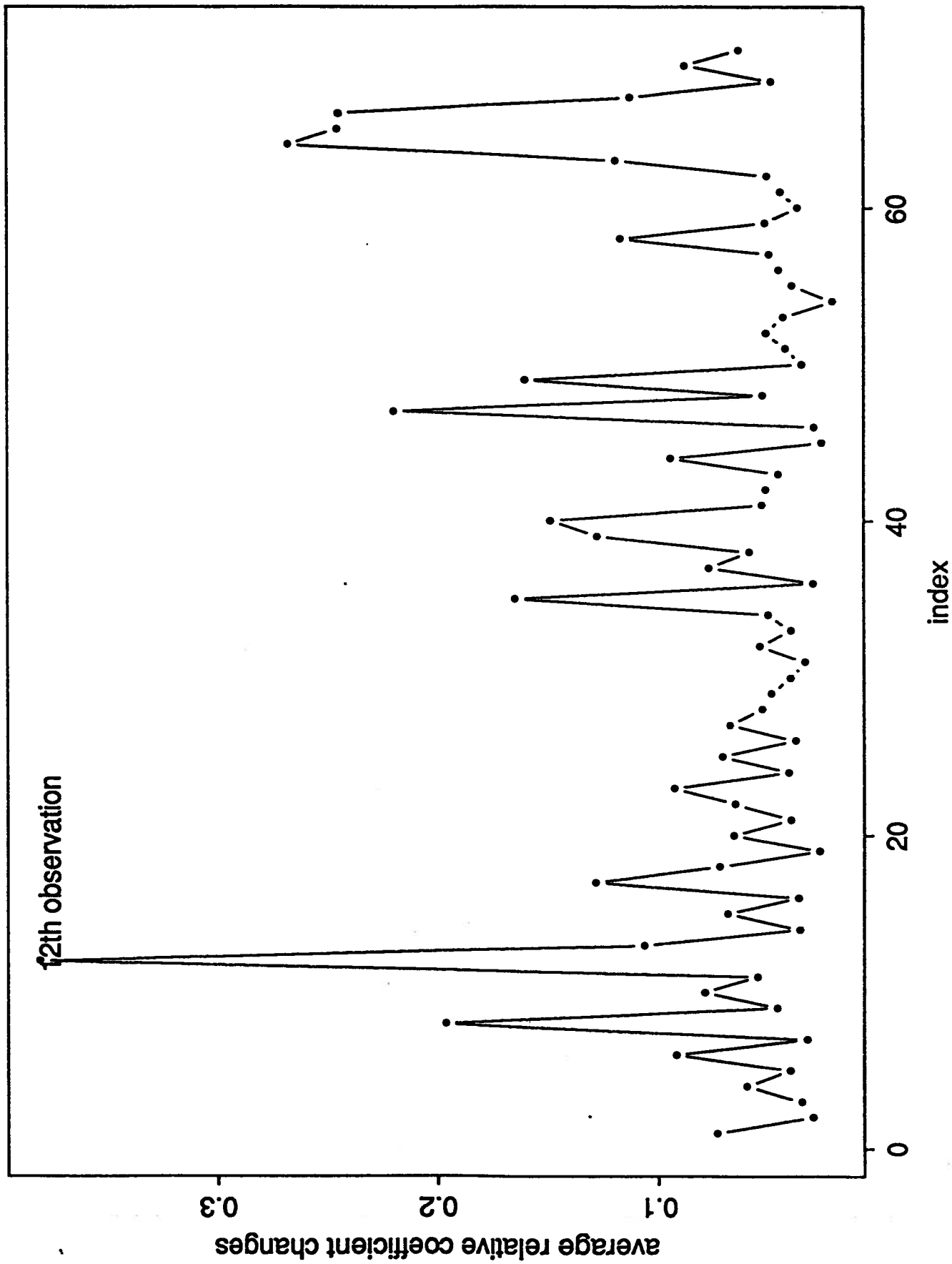

Figure 2.4: The index plot of the average relative coefficient changes from the fitted 3-component mixed Poisson regression model for the patent data.

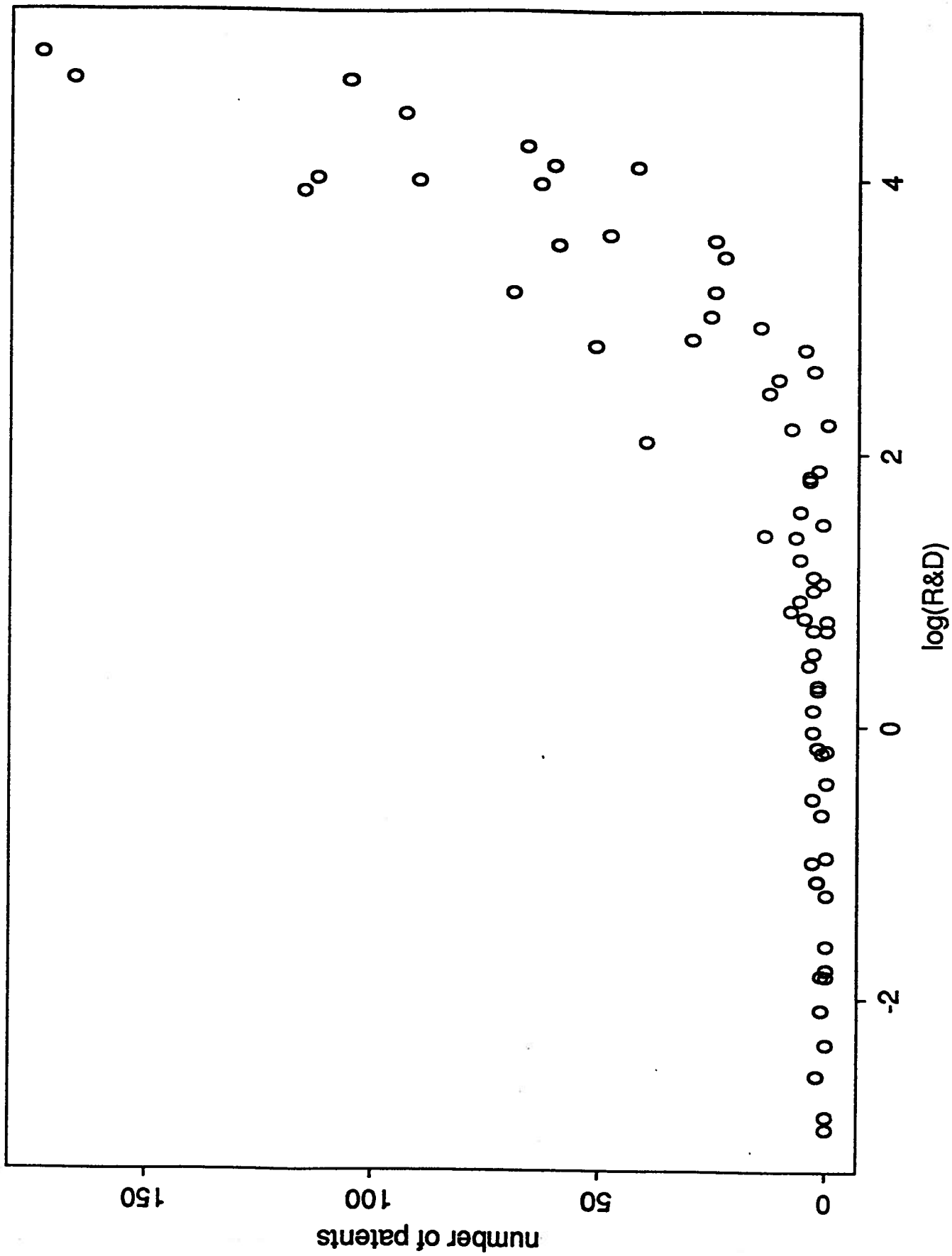

Figure 2.5: The plot of the patent data.

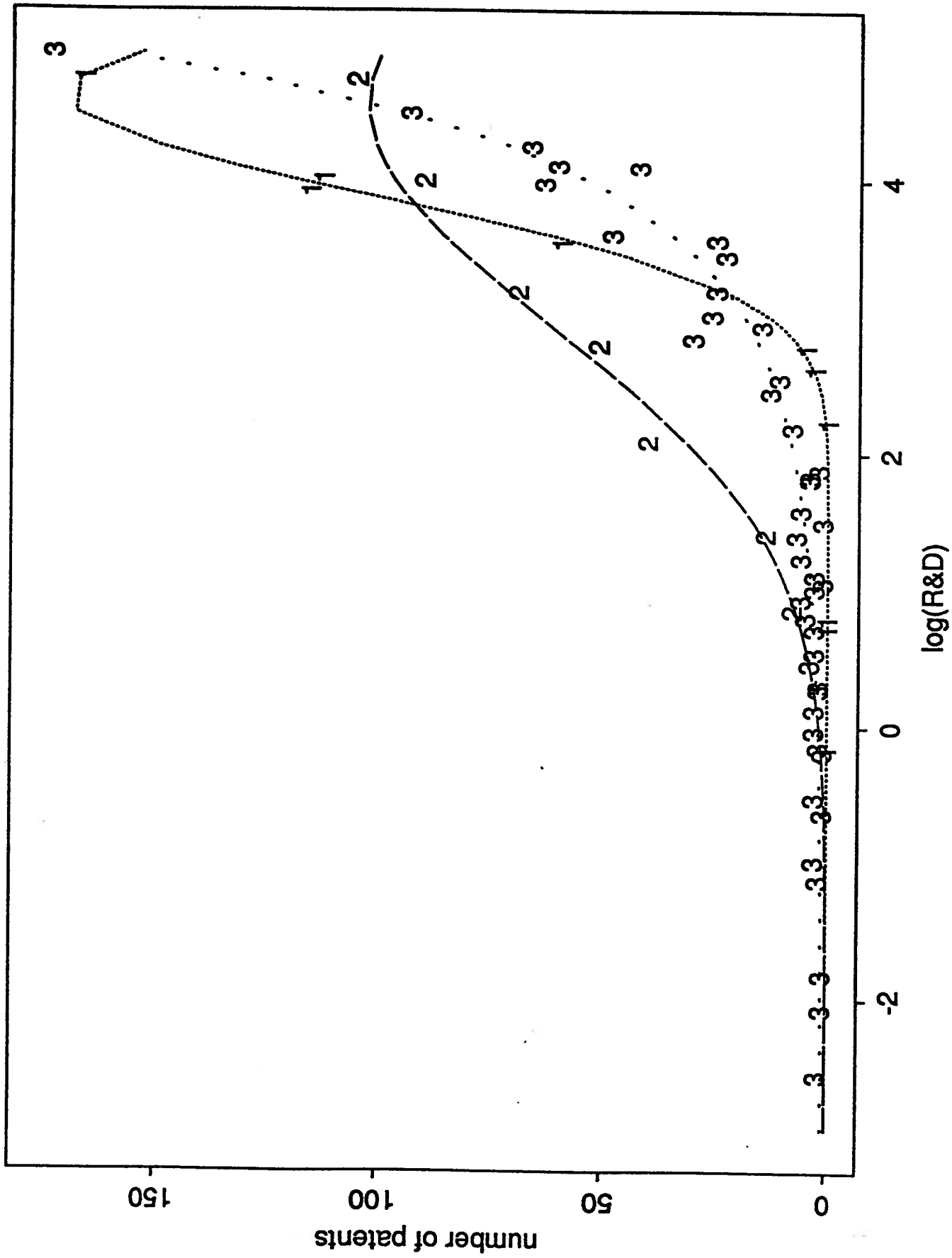

Figure 2.6: The classification of the patent data according to the estimated posterior probabilities based on the fitted mixed Poisson regression model.

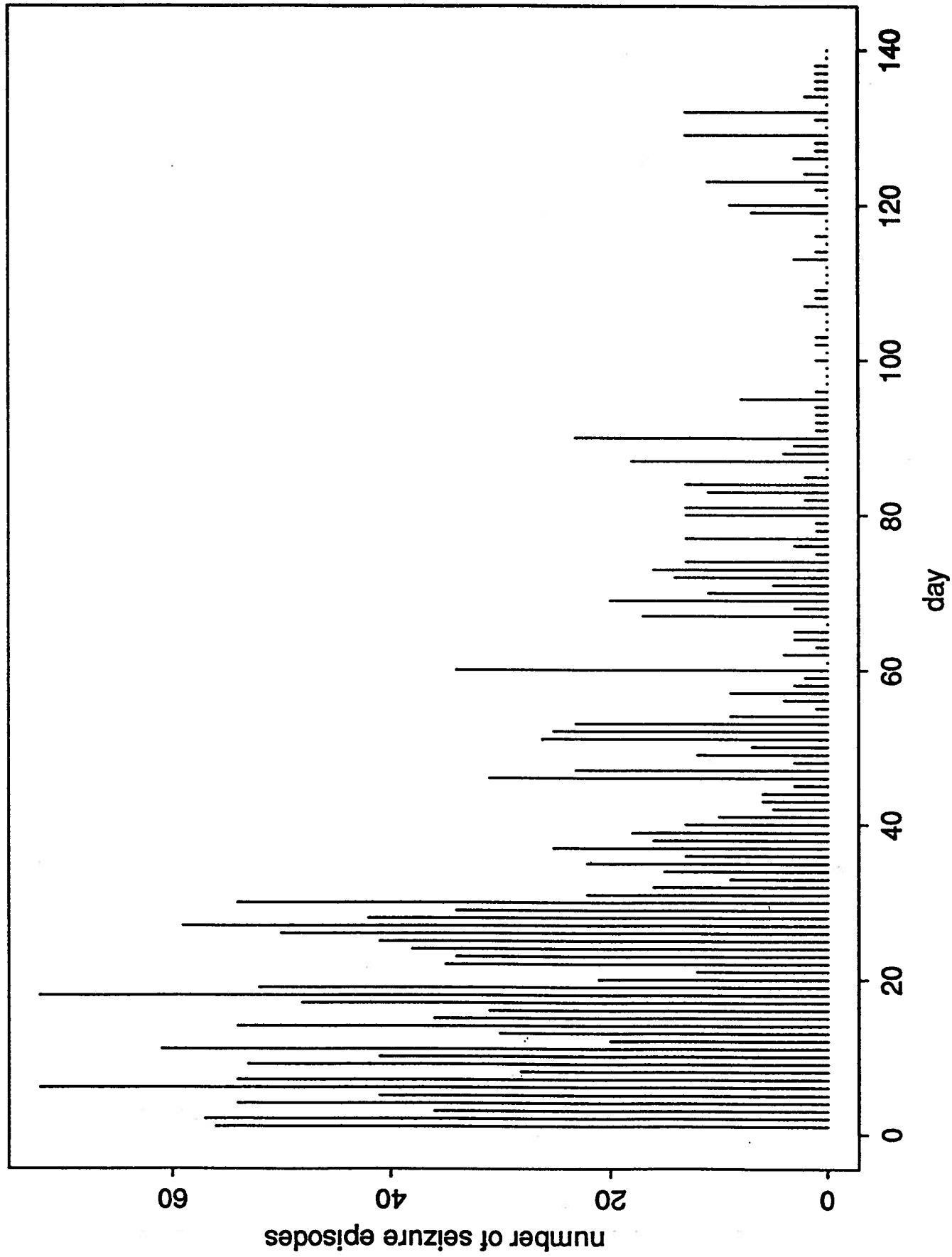

Figure 2.7: Daily epileptic seizure counts.

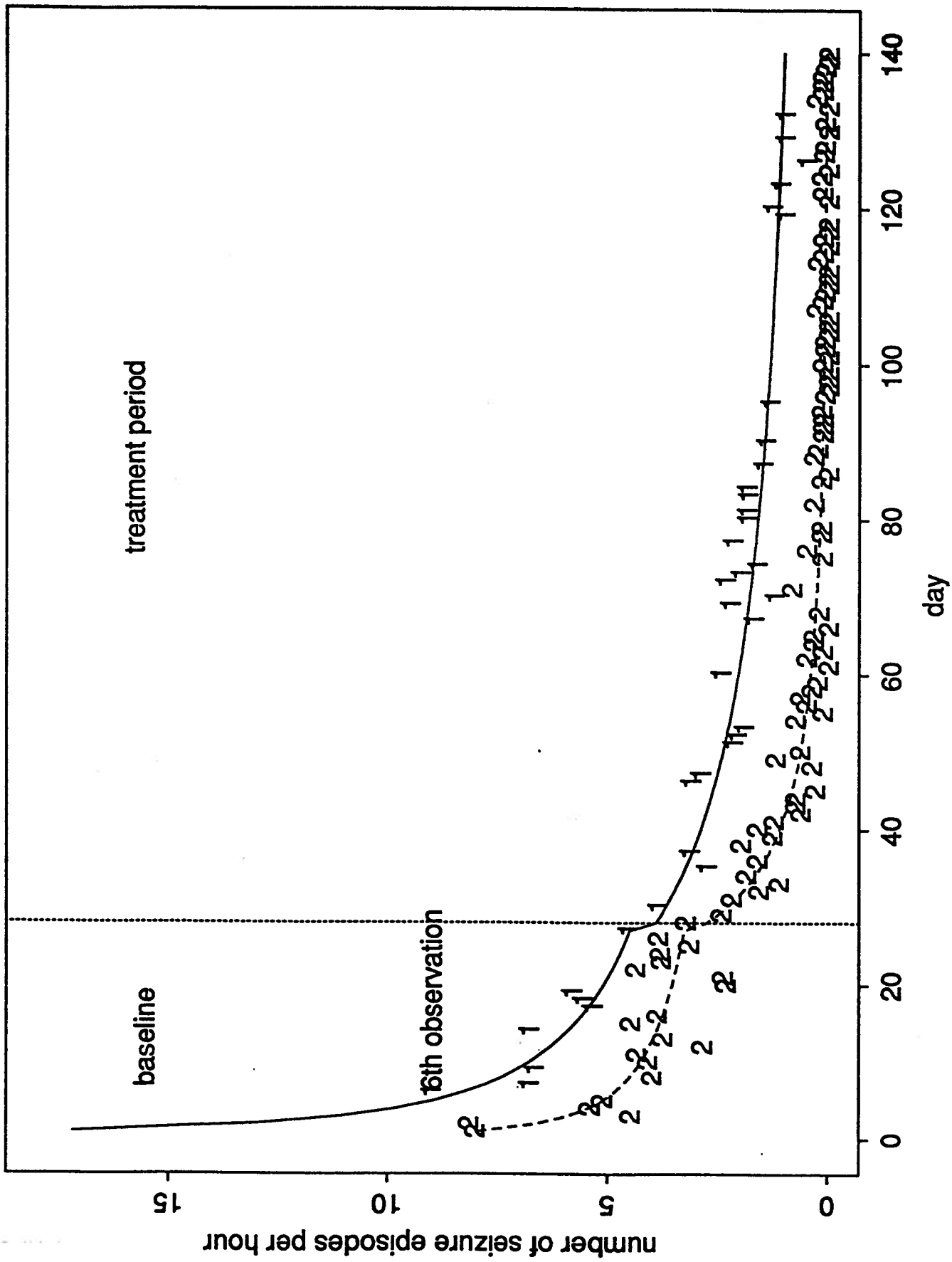

Figure 2.8: Estimated hourly seizure rates and classification of the seizure data according to the estimated posterior probabilities based on the fitted mixed Poisson regression model.

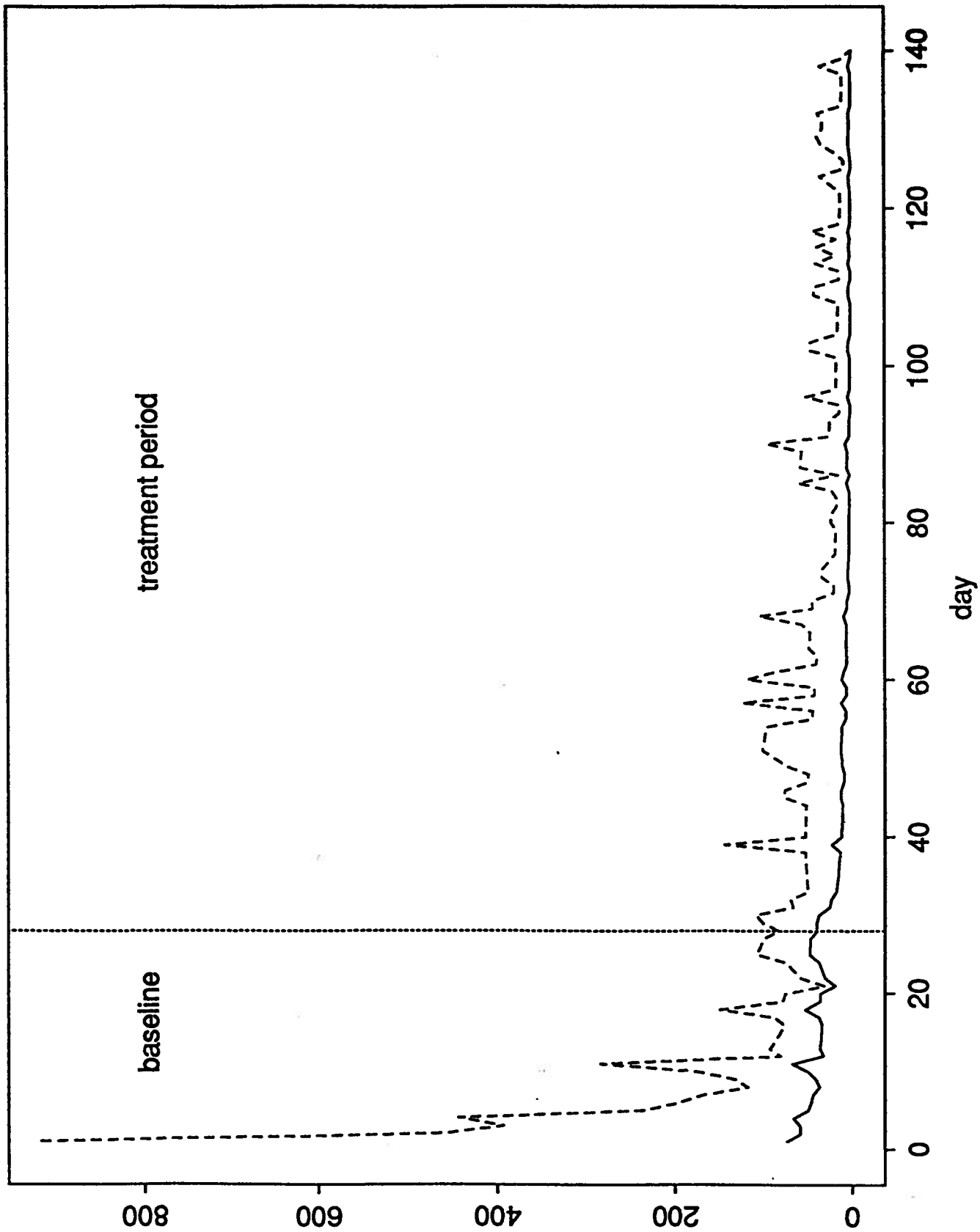

Figure 2.9: Estimated mean and variance based on the fitted mixed Poisson regression model for the seizure data.

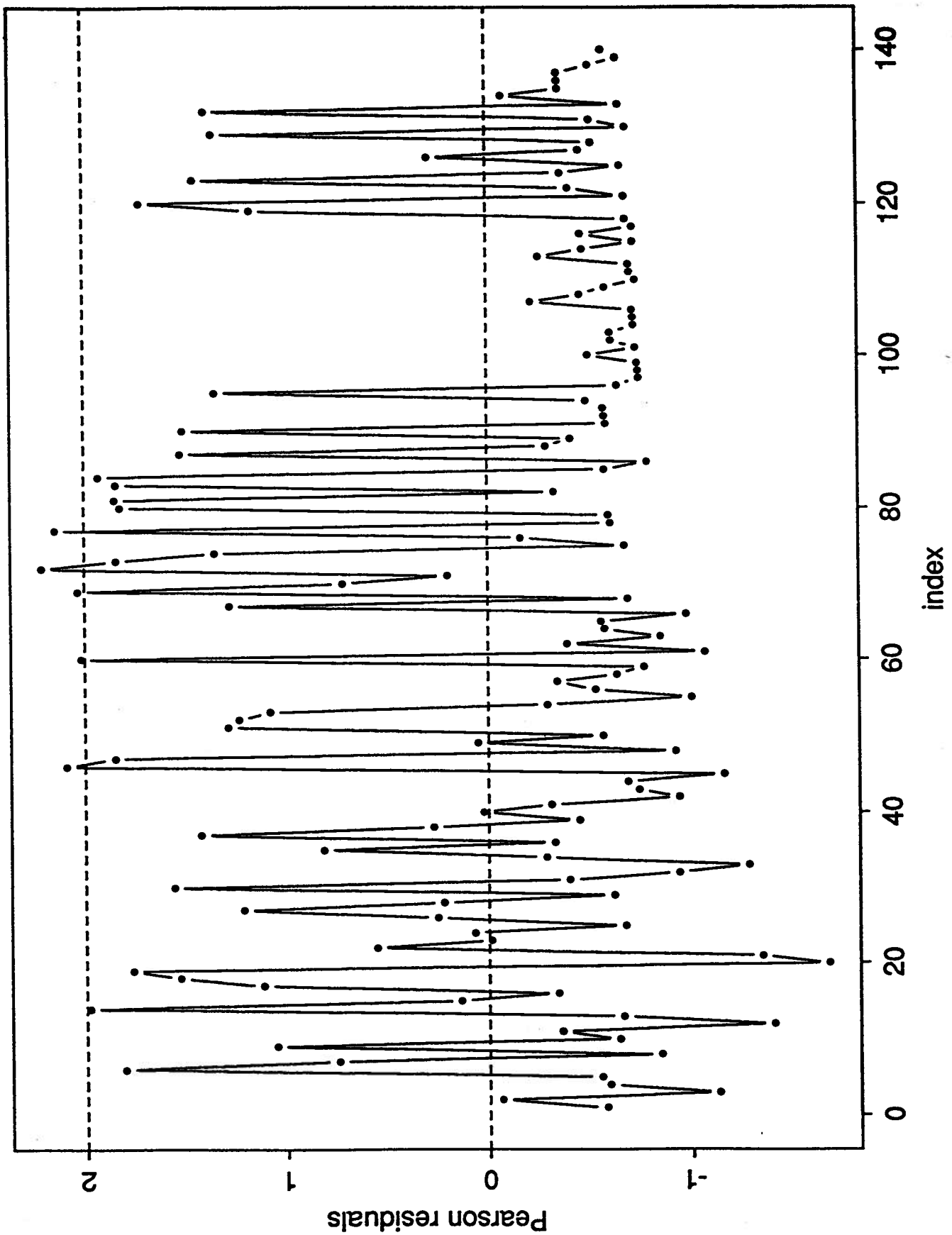

Figure 2.10: The index plot of the Pearson residuals from the fitted mixed Poisson regression model for the seizure data.

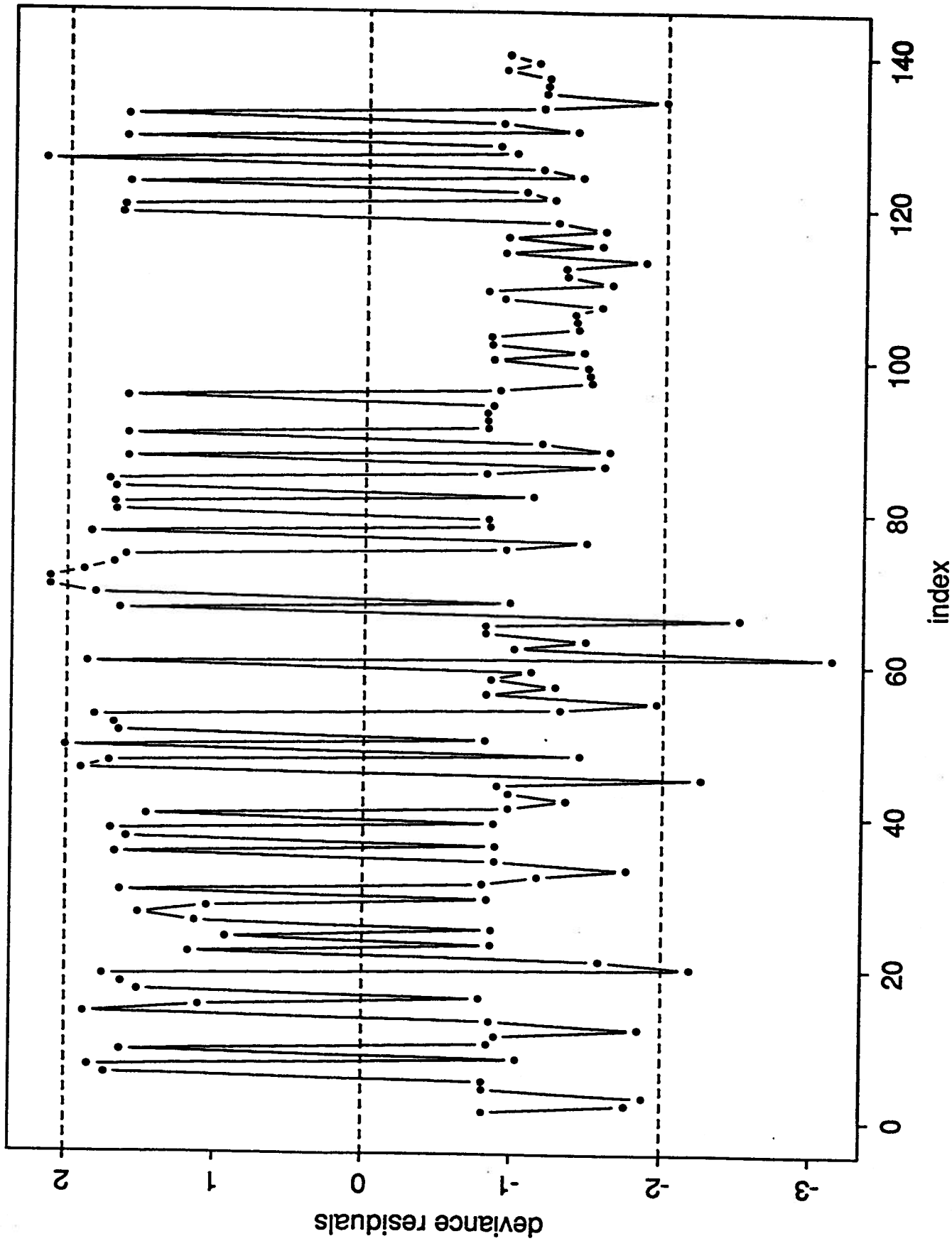

Figure 2.11: The index plot of the deviance residuals from the fitted mixed Poisson regression model for the seizure data.

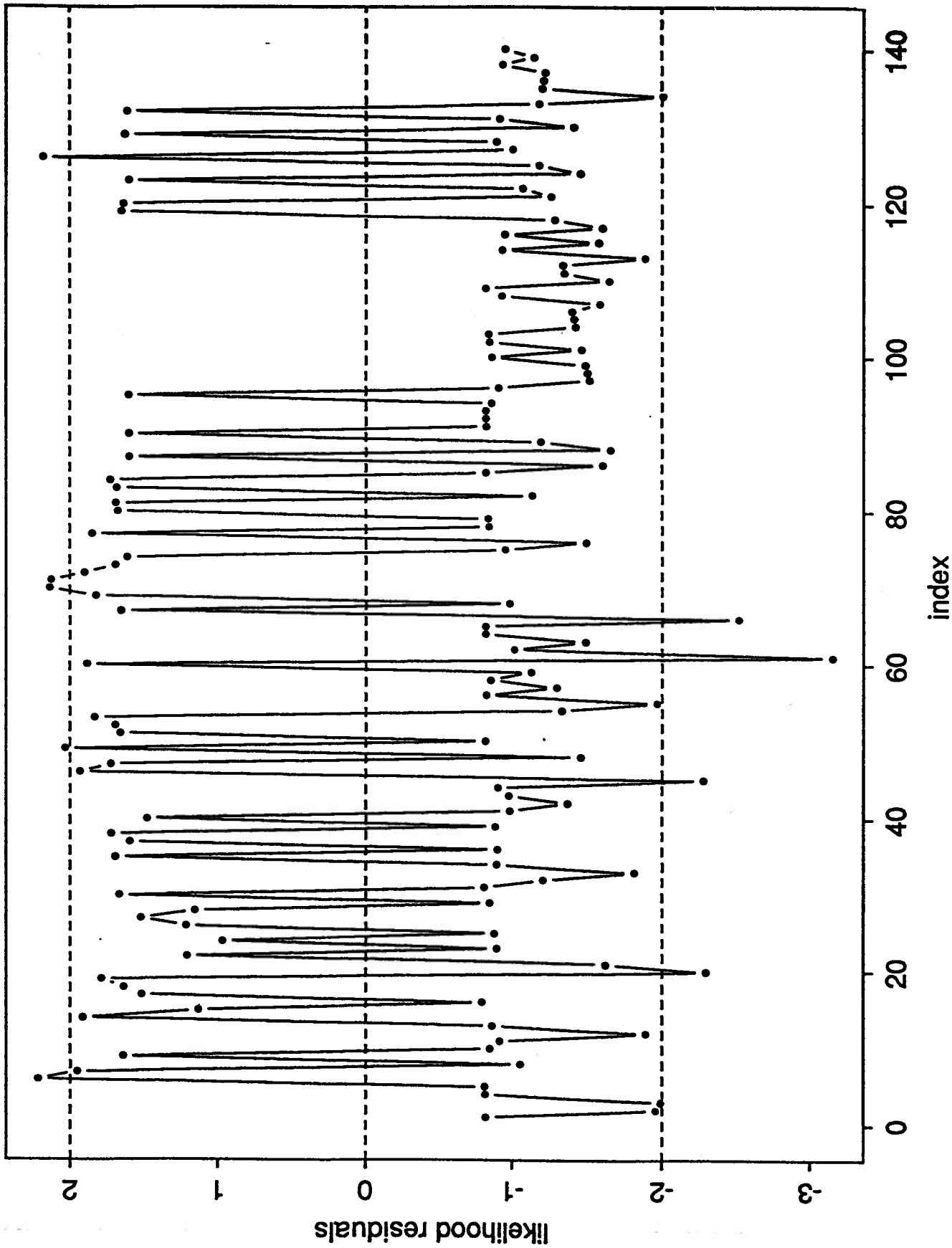

Figure 2.12: The index plot of the likelihood residuals from the fitted mixed Poisson regression model for the seizure data.

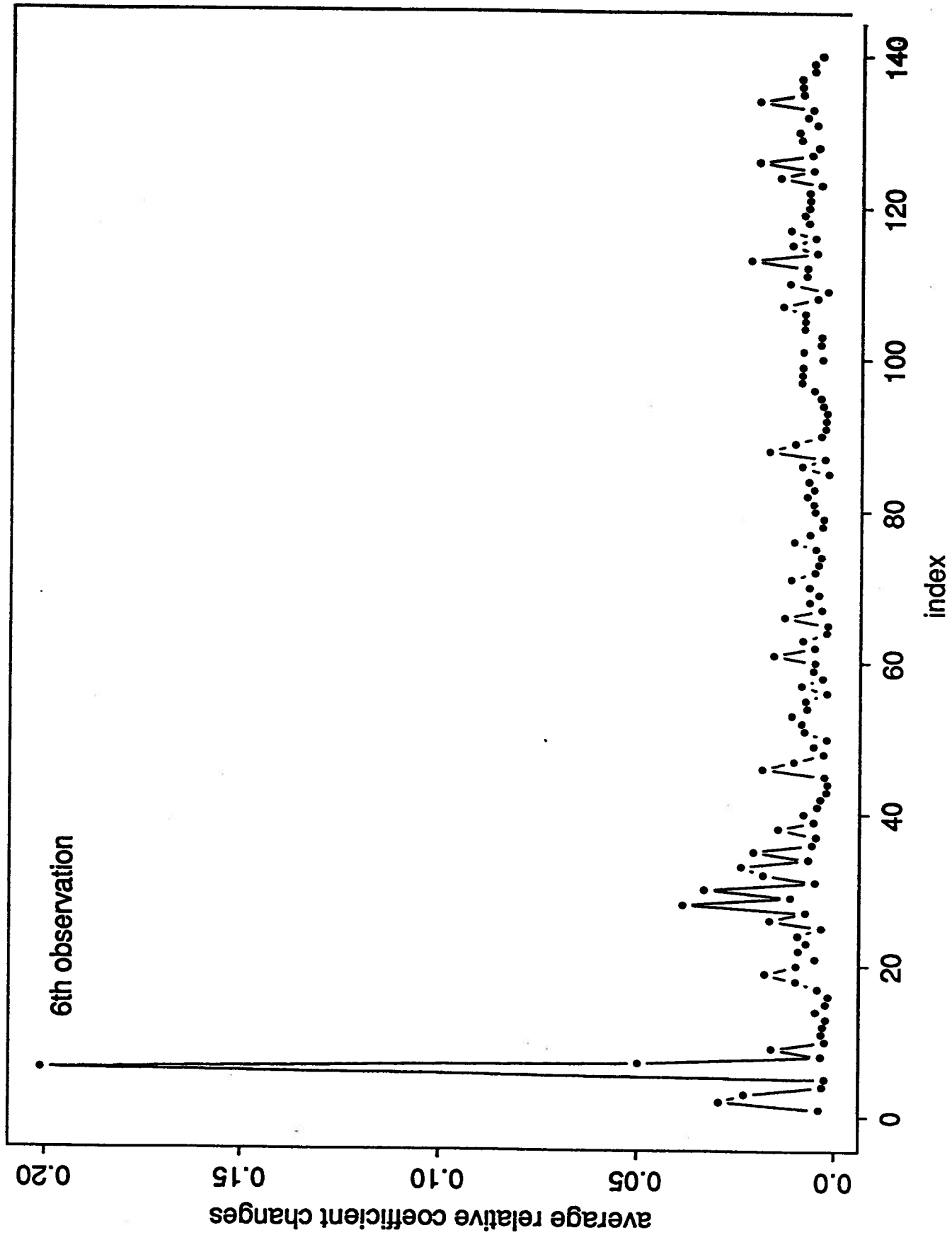

Figure 2.13: The index plot of the average relative coefficient changes from the fitted mixed Poisson regression model for the seizure data.

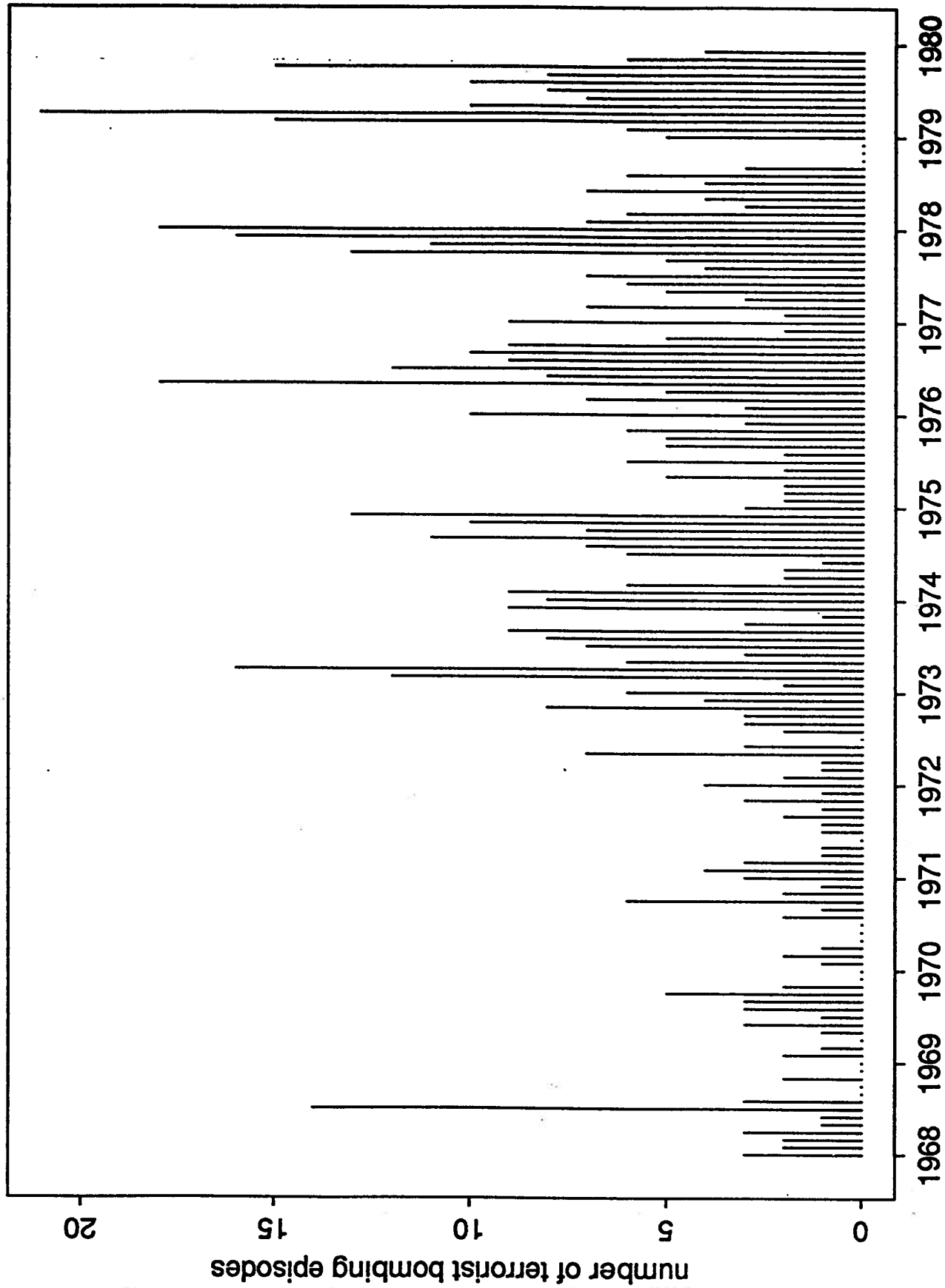

Figure 2.14: The time plot of the terrorist bombing data.

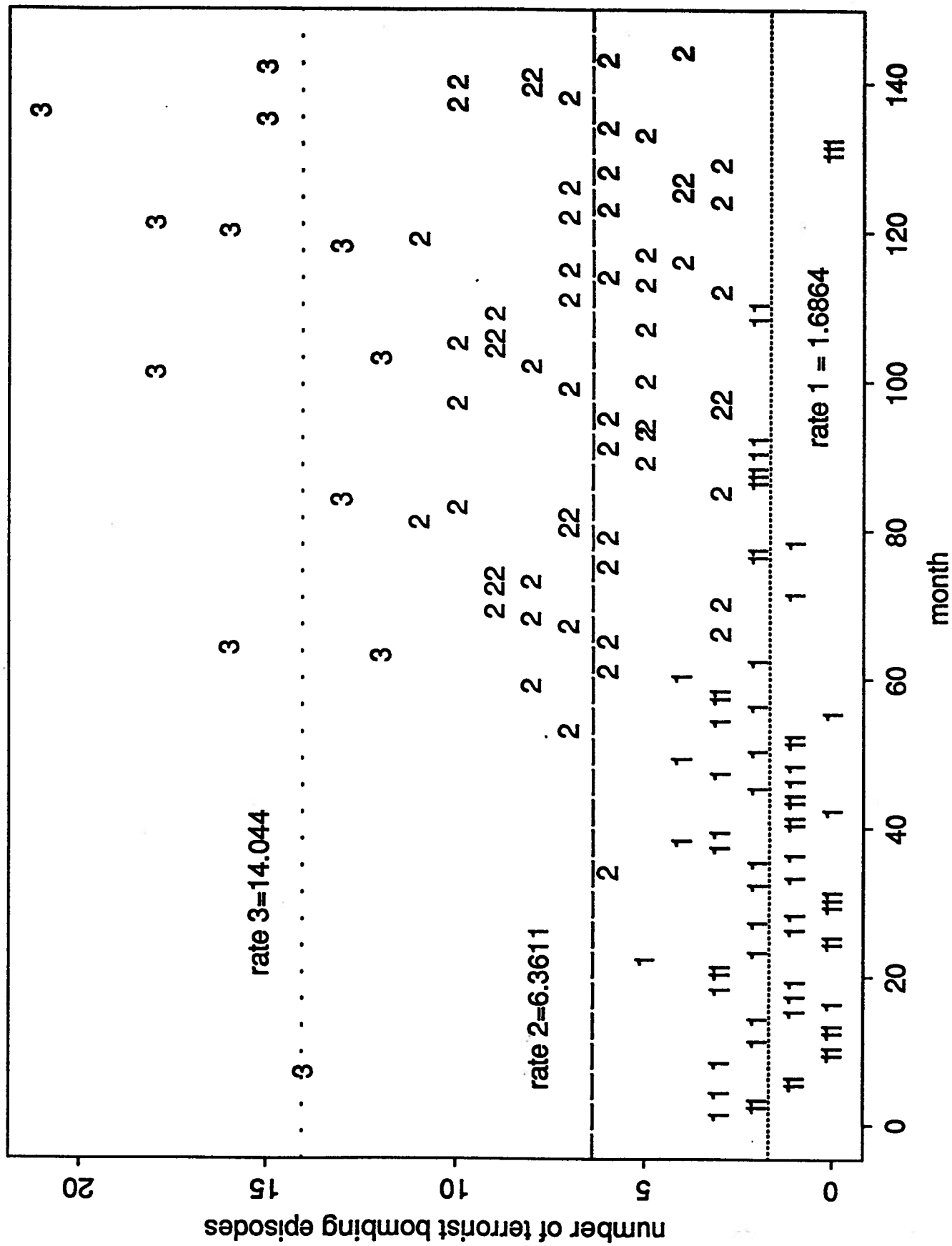

Figure 2.15: Classification of the terrorist bombing episodes according to the estimated posterior probabilities based on the fitted mixed Poisson regression model.

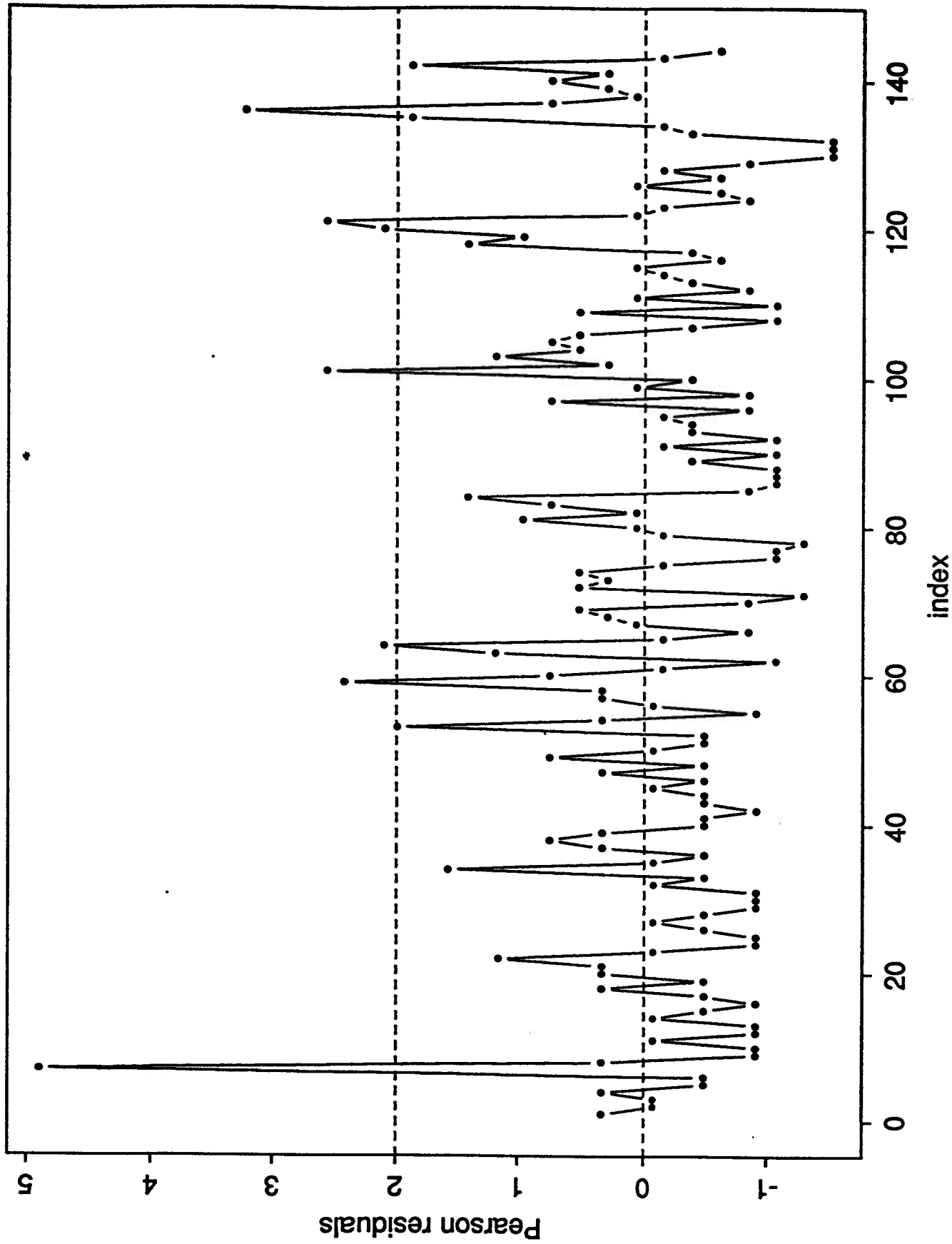

Figure 2.16: The index plot of the Pearson residuals from the fitted mixed Poisson regression model for the terrorist bombing data.

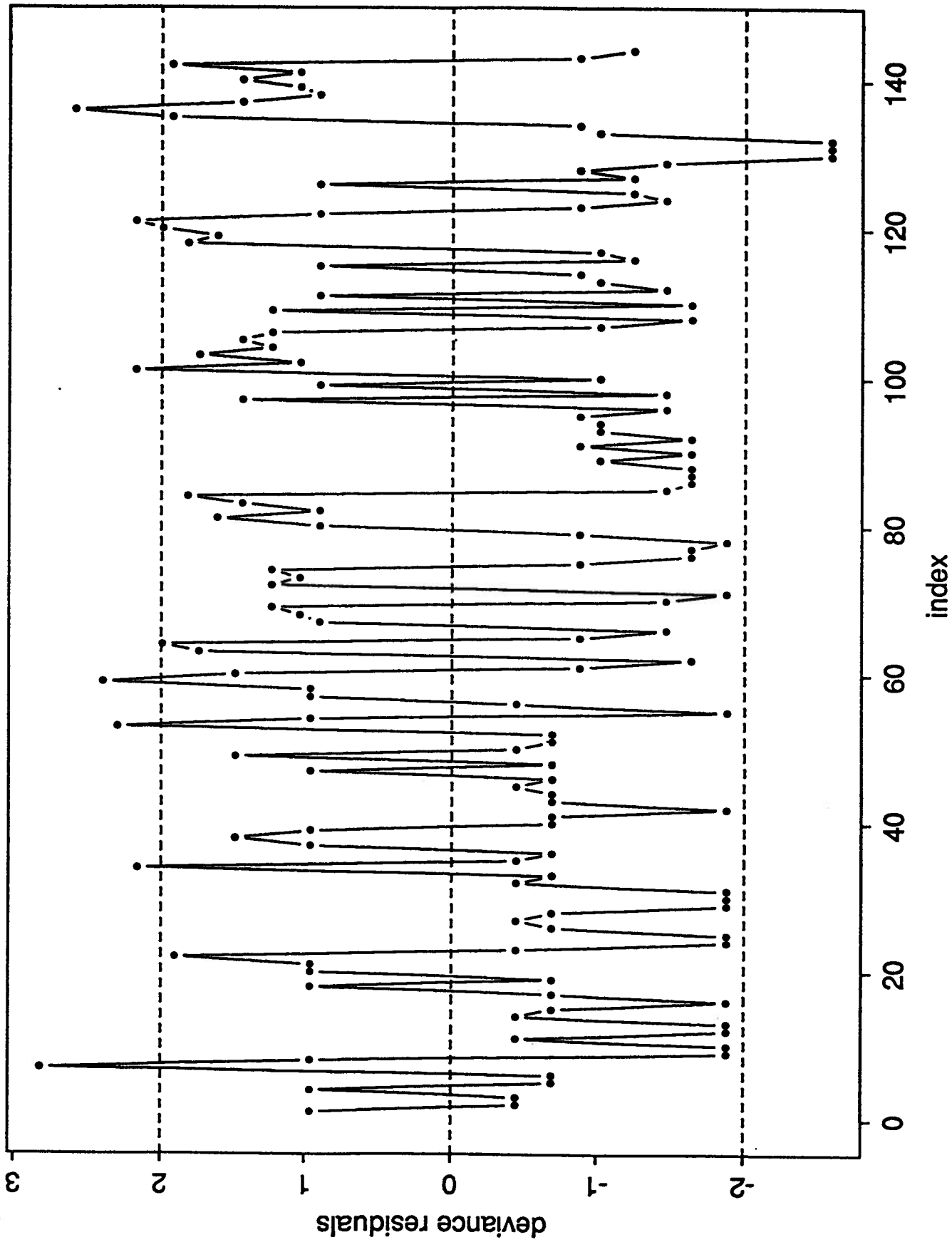

Figure 2.17: The index plot of the deviance residuals from the fitted mixed Poisson regression model for the terrorist bombing data.

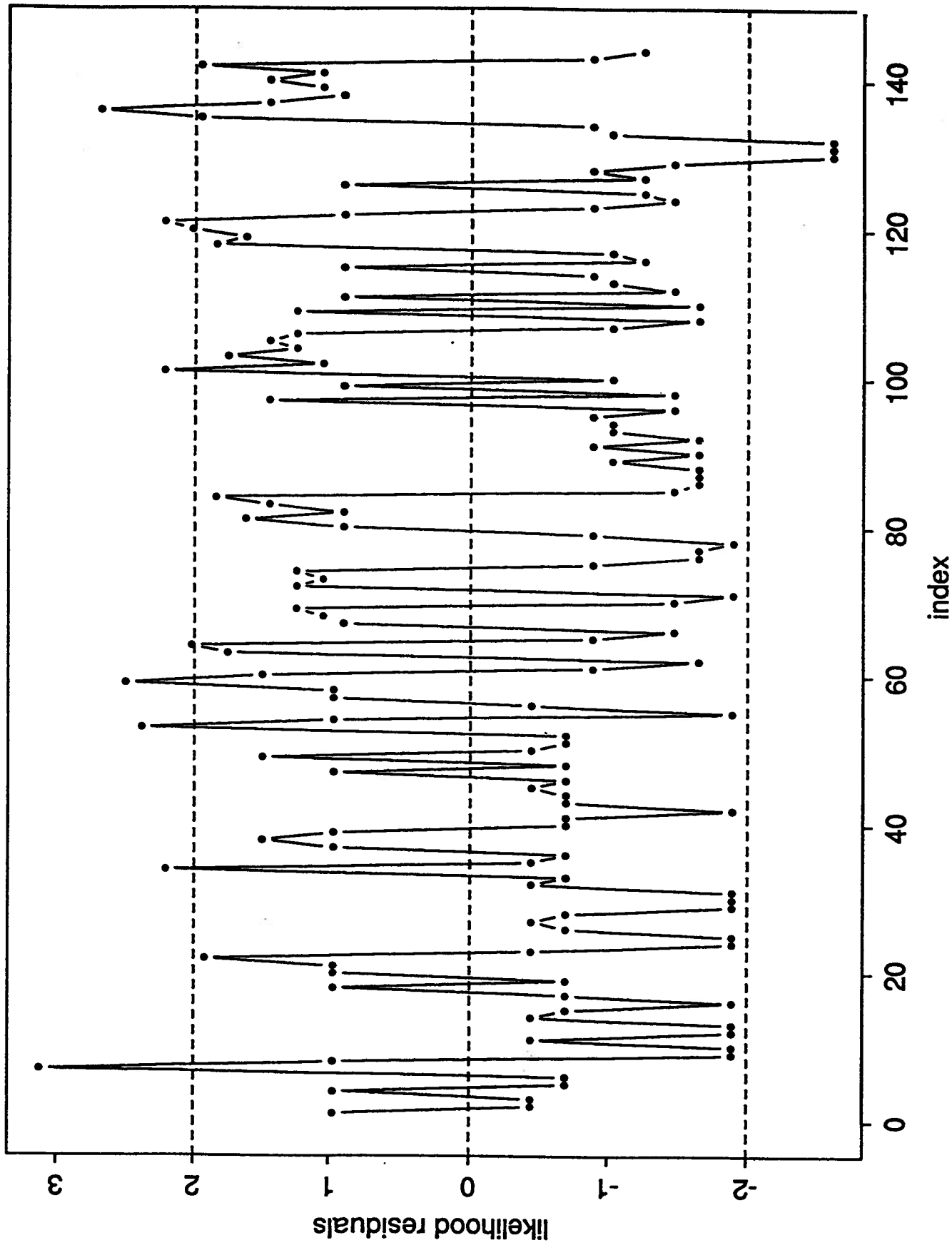

Figure 2.18: The index plot of the likelihood residuals from the fitted mixed Poisson regression model for the terrorist bombing data.

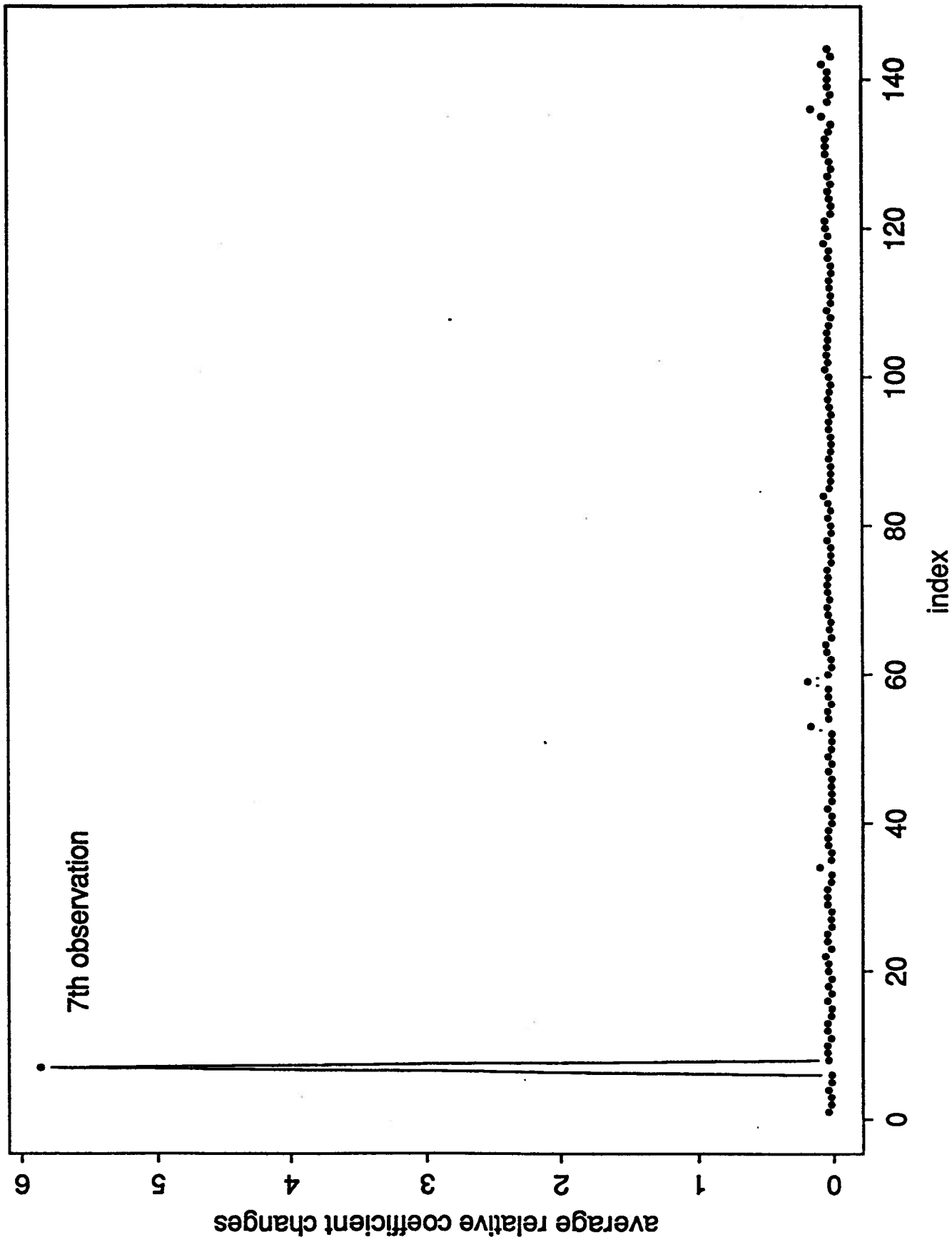

Figure 2.19: The index plot of the average relative coefficient changes from the fitted mixed Poisson regression for the terrorist bombing data.

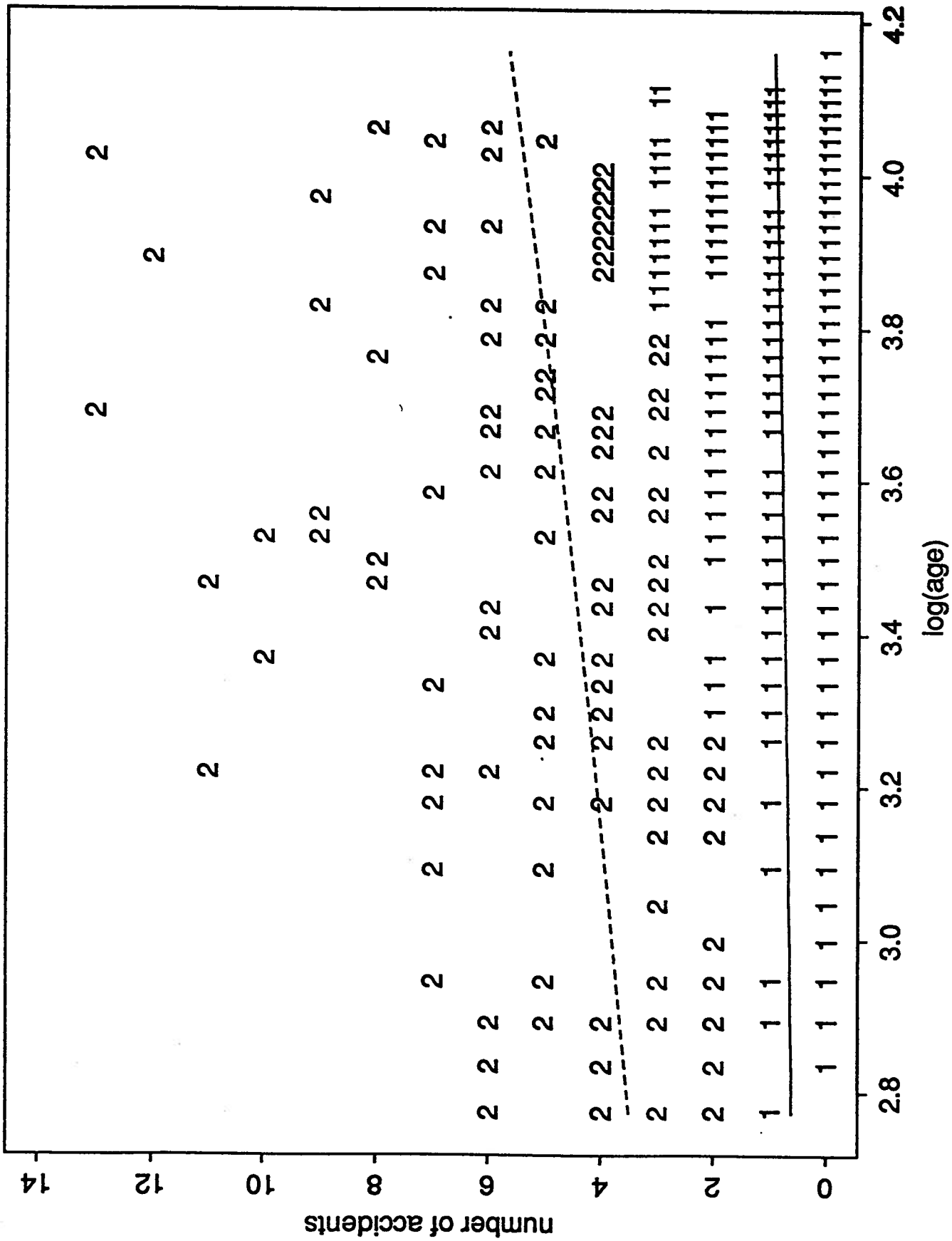

Figure 2.20: Classification of the accident data according to the estimated posterior probabilities based on the fitted mixed Poisson regression model.

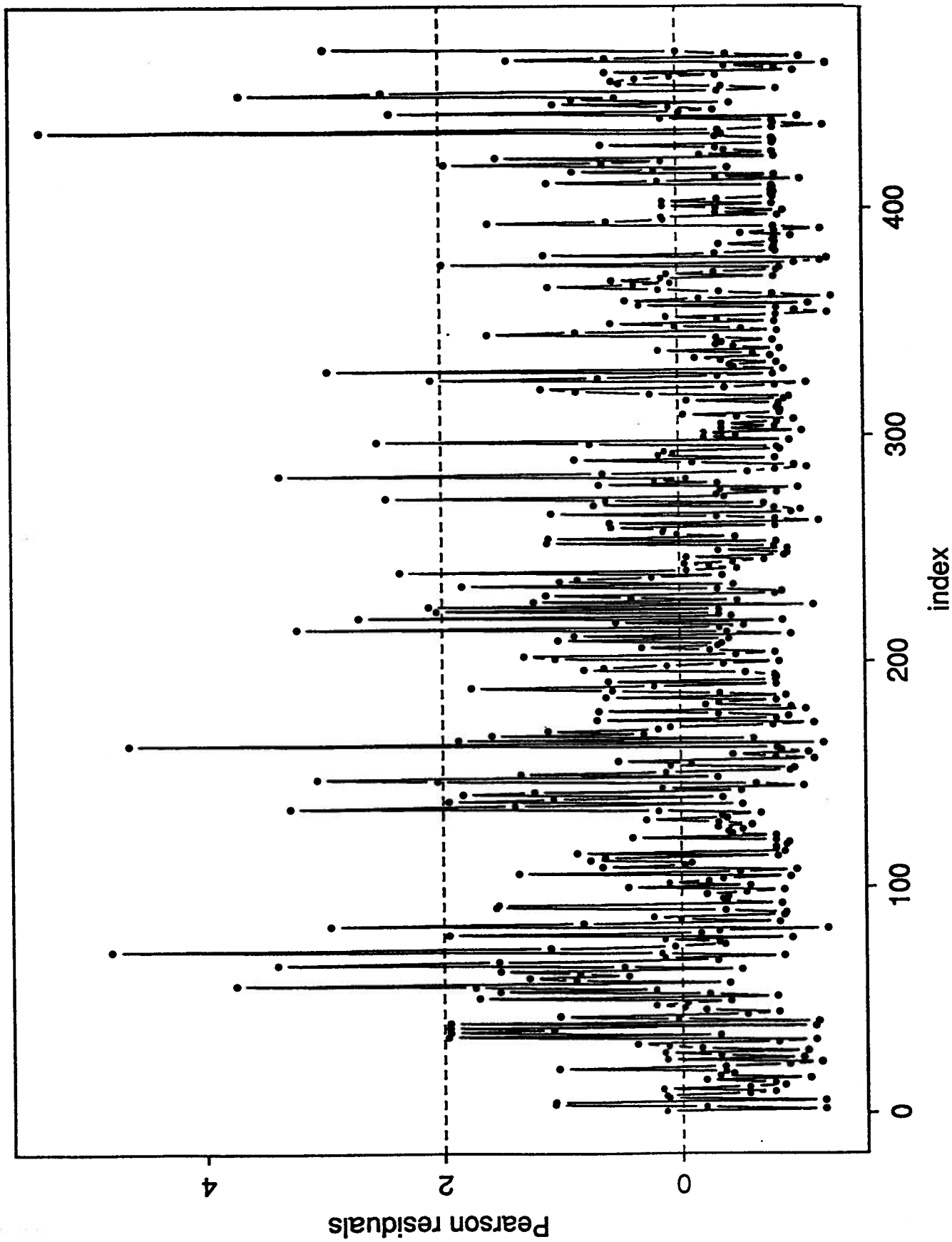

Figure 2.21: The index plot of the Pearson residuals from the fitted mixed Poisson regression model for the accident data.

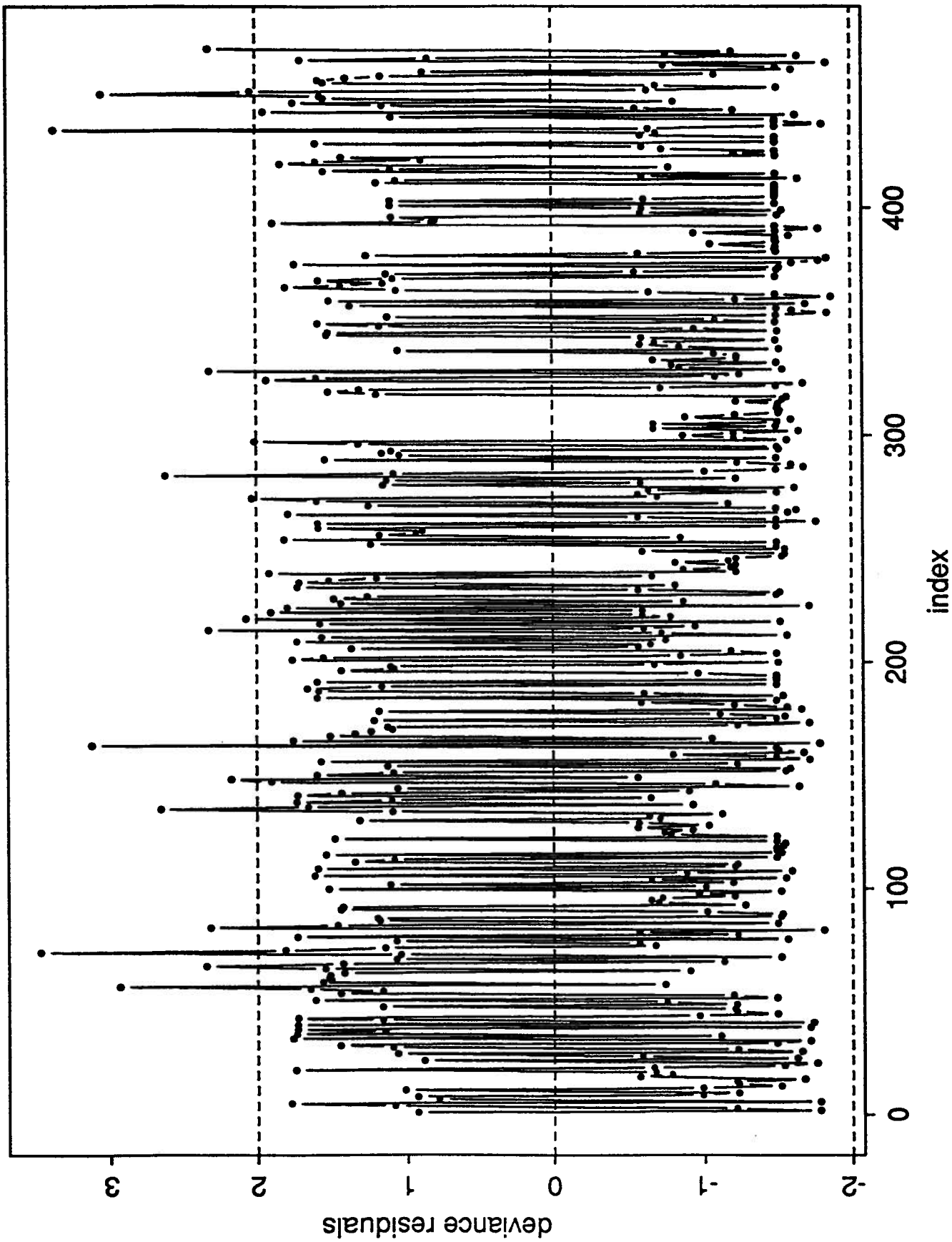

Figure 2.22: The index plot of the deviance residuals from the fitted Poisson regression model for the accident data.

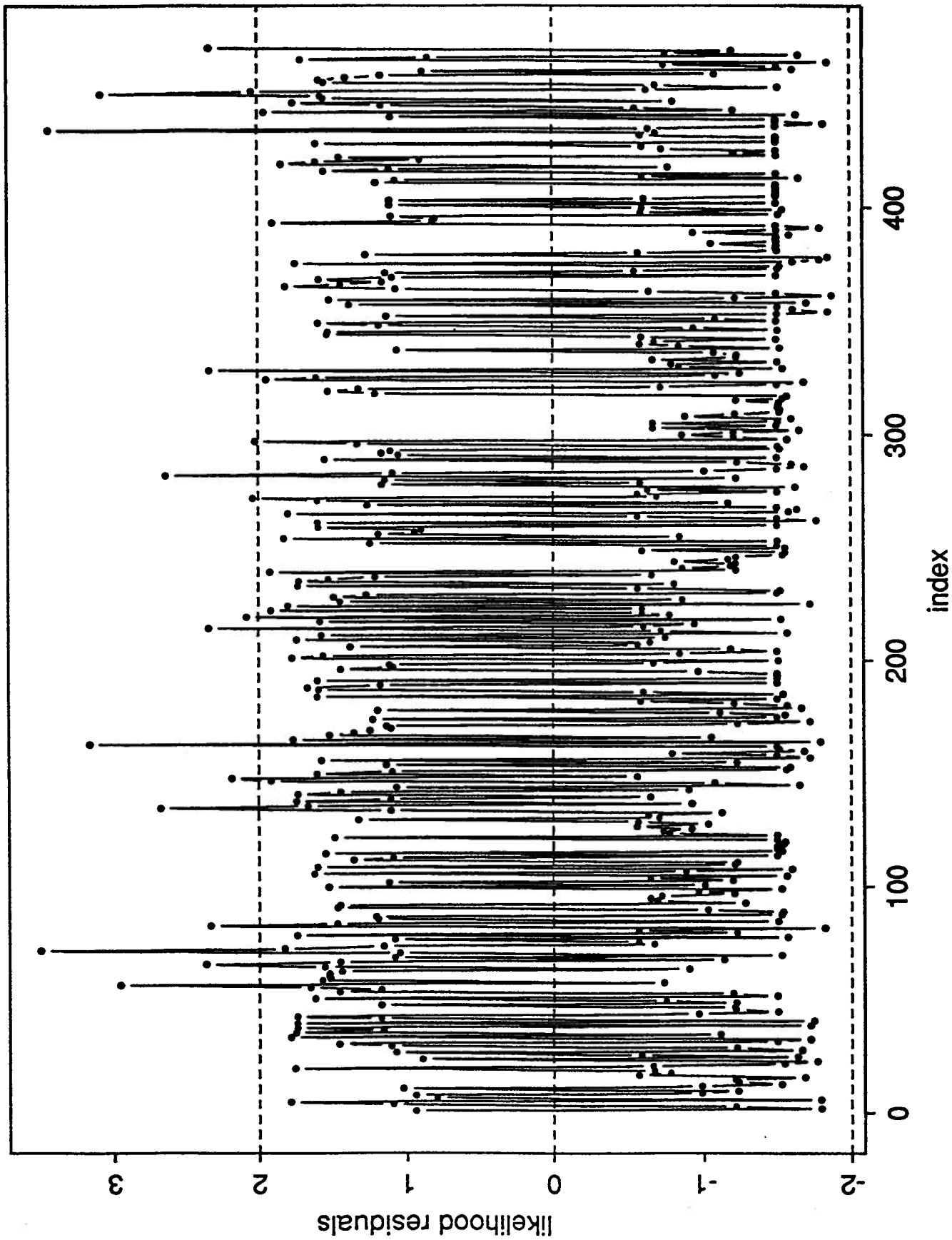

Figure 2.23: The index plot of the likelihood residuals from the fitted mixed Poisson regression model for the accident data.

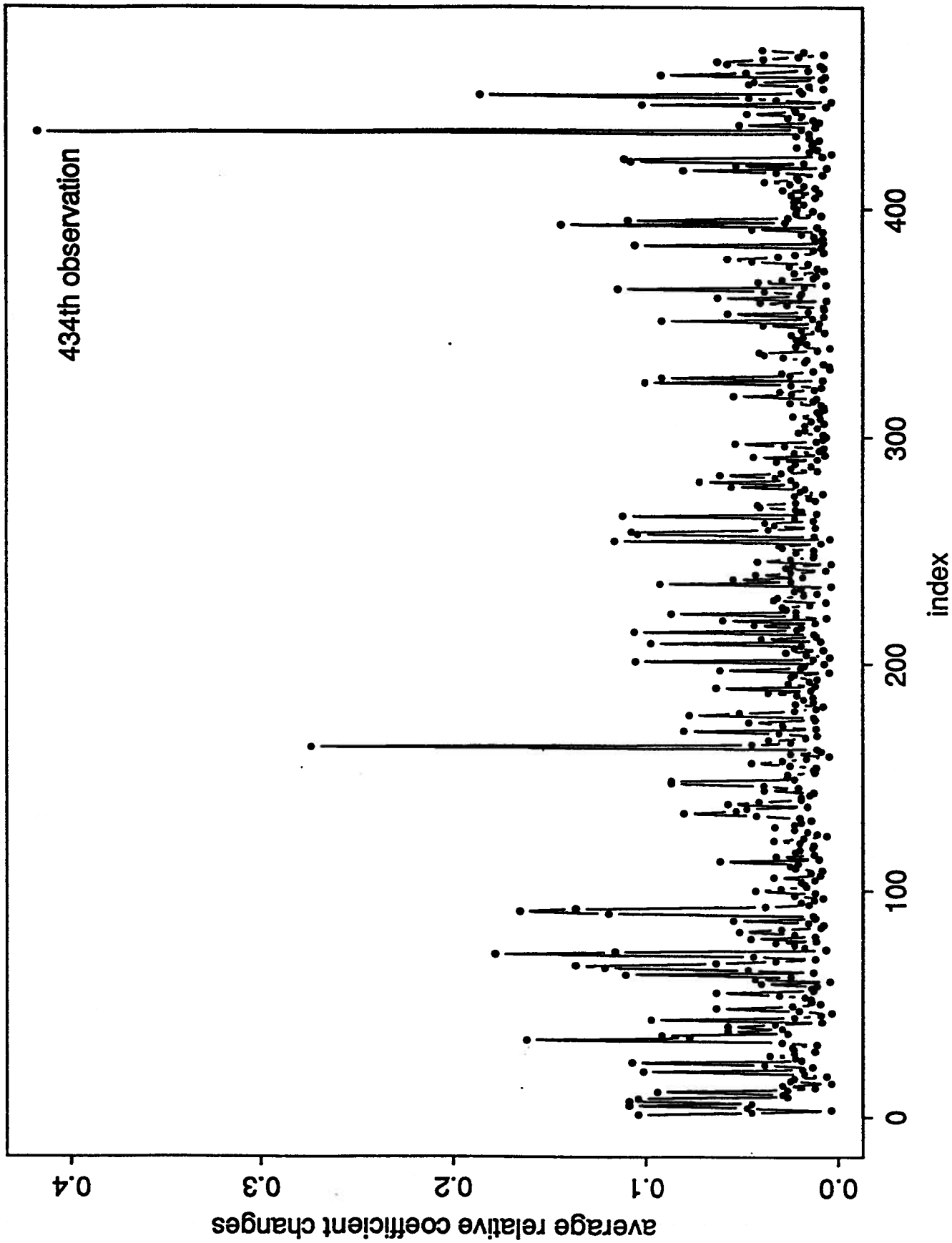

Figure 2.24: The index plot of the average relative coefficient changes from the fitted mixed Poisson regression model for the accident data.

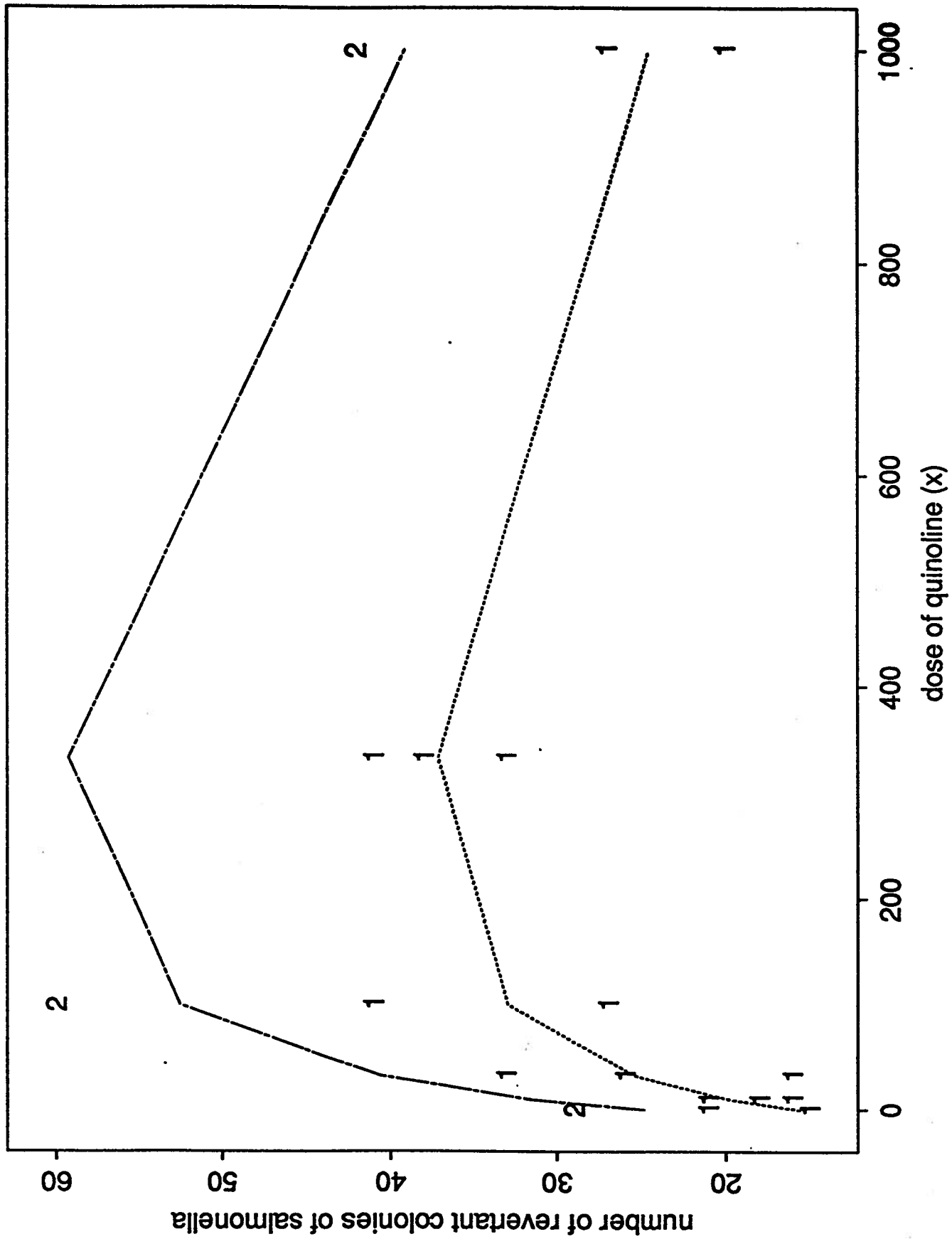

Figure 2.25: Classification of the Ames salmonella assay data according to the estimated posterior probabilities based on the fitted mixed Poisson regression model.

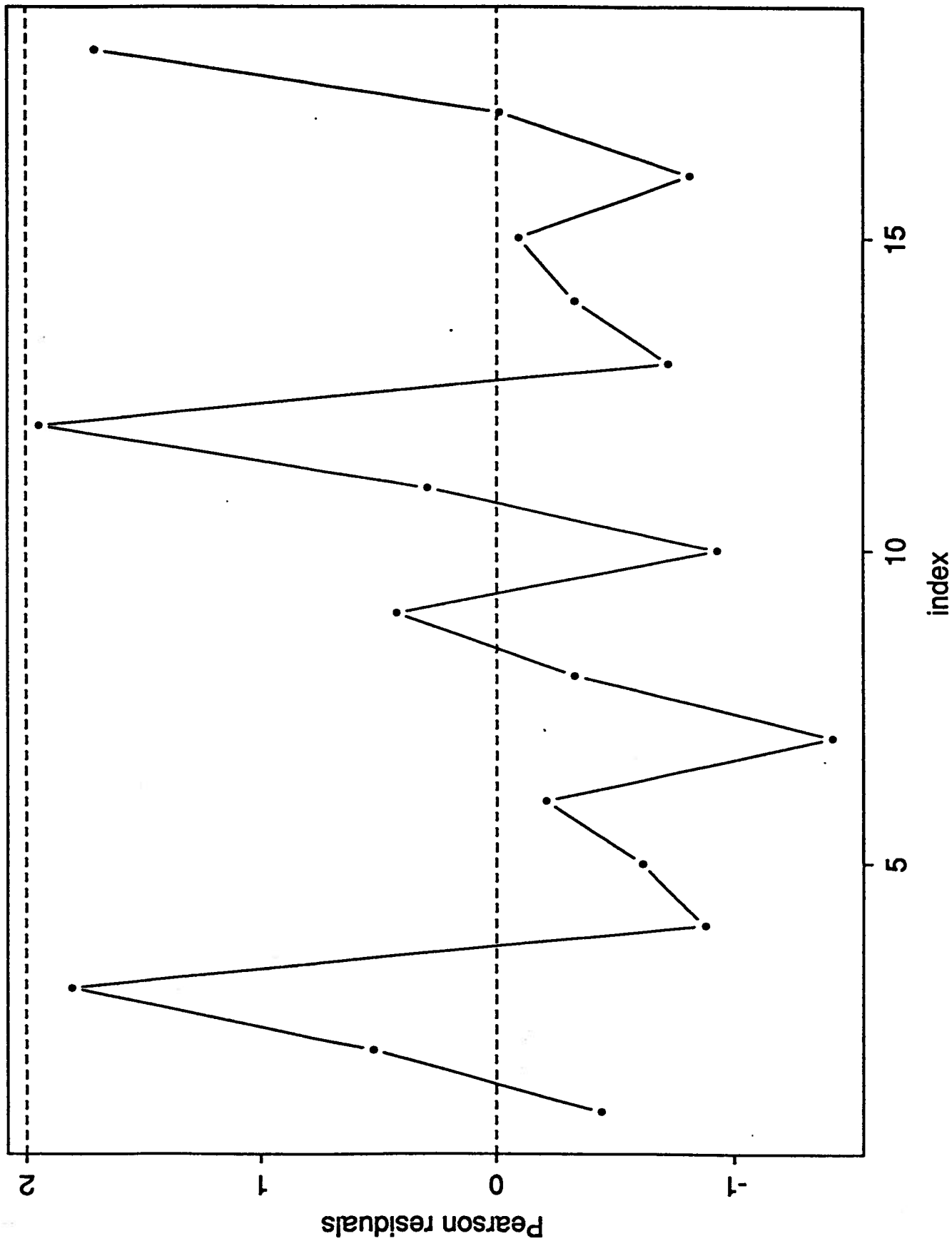

Figure 2.26: The index plot of the Pearson residuals from the fitted mixed Poisson regression model for the Ames salmonella assay data.

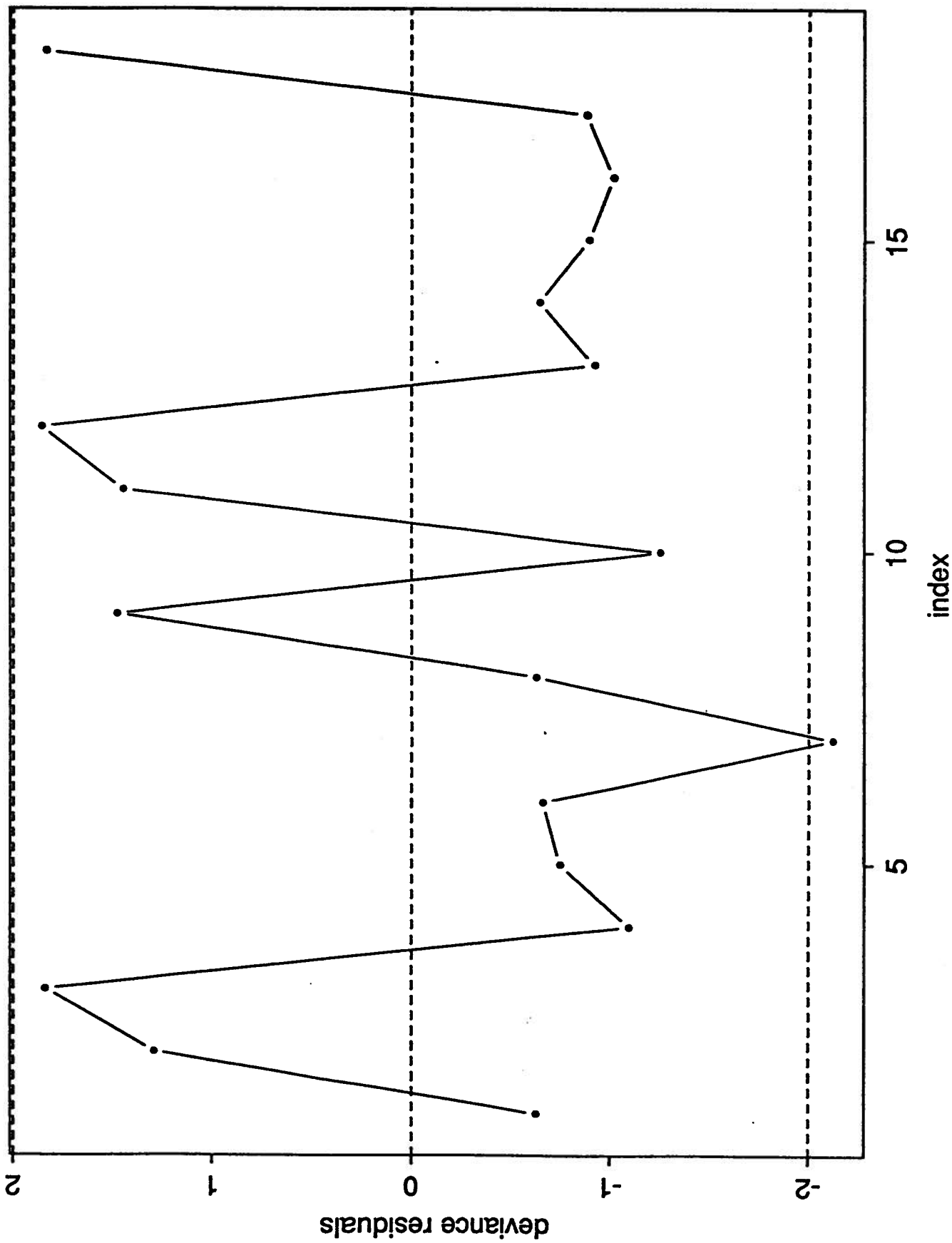

Figure 2.7: The index plot of the deviance residuals from the fitted mixed Poisson regression model for the Ames salmonella assay data.

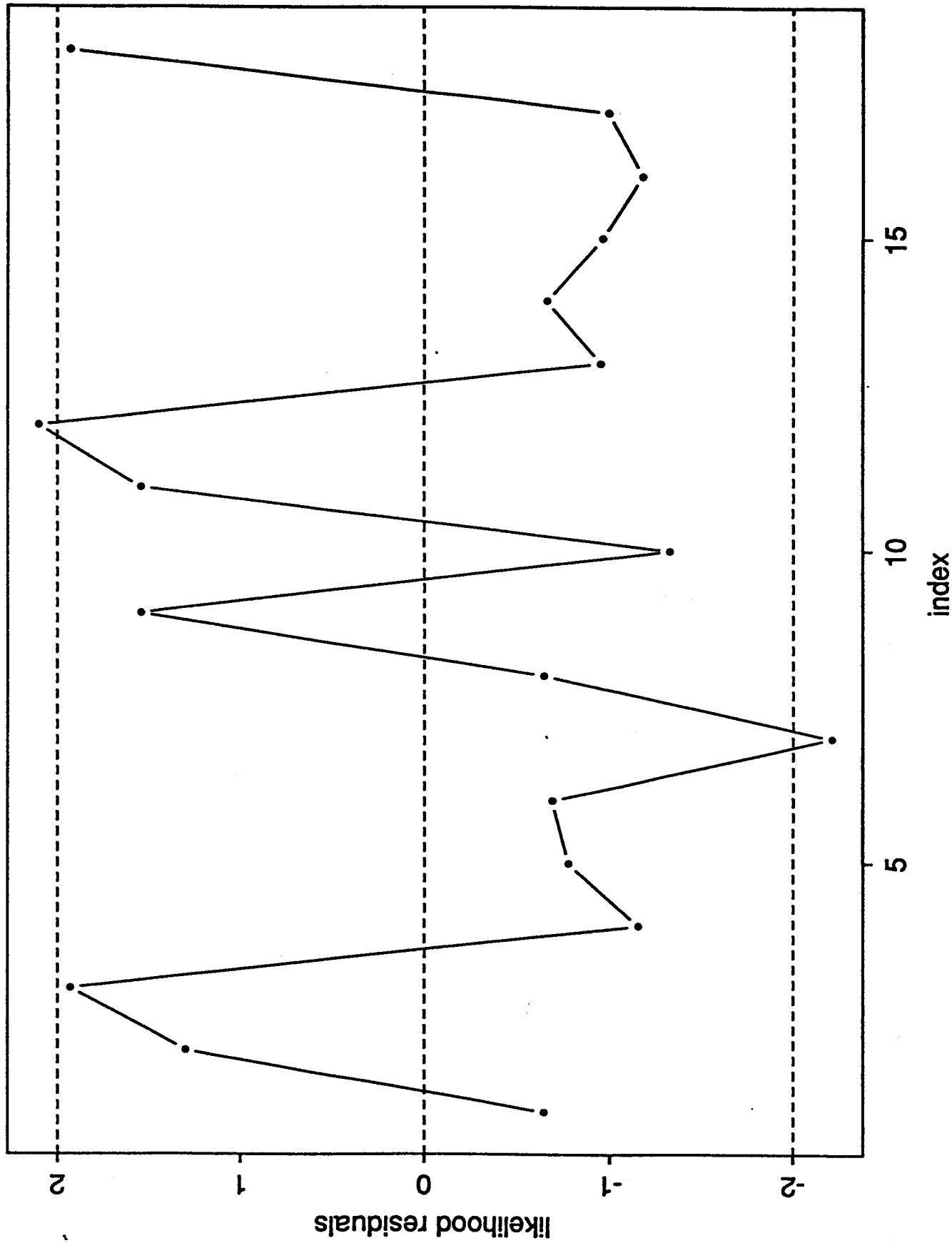

Figure 2.28: The index plot of the likelihood residuals from the fitted mixed Poisson regression model for the Ames salmonella assay data.

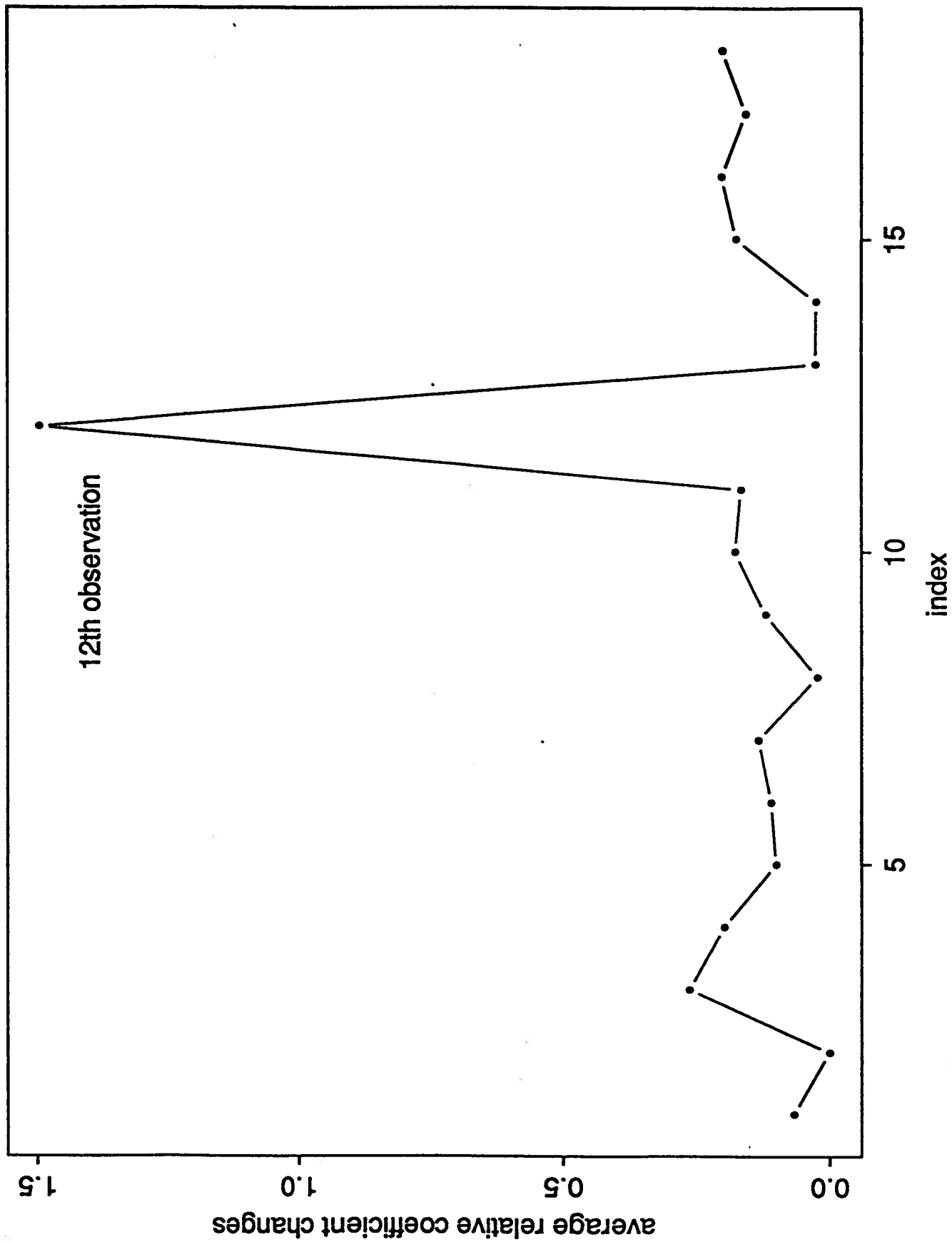

Figure 2.29: The index plot of the average relative coefficient changes from the mixed Poisson regression model for the Ames salmonella assay data.

Chapter 3

Mixed Logistic Regression Models

3.1 Logistic Regression and Its Modifications

The logistic regression model has been widely used for analyzing count 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 the model fits poorly, suggesting the need for alternative models. In this case, it is not uncommon that observed data are overdispersed in terms of the binomial assumption. In the second part of this dissertation, mixed logistic regression models are introduced and investigated. These models are applicable in several different situations where the usual logistic regression model is inadequate. They provide an alternative way to quasi-likelihood approach and others for modelling extra-binomial variation with a more meaningful interpretation.

Suppose that the ith response Y_i is a count of successes in m_i trials, and associated with this response is a covariate vector $x_i = (x_{i1}, \ldots, x_{ir})'$ for $1 \leq i \leq n$. The logistic regression model assumes that the Y_i are distributed independently binomial(m_i, π_i) with density function given by

$$f(y_i \mid \alpha, x_i, m_i) = \begin{pmatrix} m_i \\ y_i \end{pmatrix} \pi_i^{y_i}(1-\pi_i)^{m_i - y_i}$$

where $\pi_i \equiv \pi(x_i, \alpha) = \exp(x_i'\alpha)/(1+\exp(x_i'\alpha))$, $\alpha \in R^r$, is a unknown regression parameter vector, m_i is an integer and $y_i = 1, \ldots, m_i$. Note that the binomial parameter π_i is

related to the linear part, $x_i'\alpha$, through a logit transformation. Note also that m_i may vary with i.

The logistic regression model may be used as follows. Sometimes, inference concerning the αs is of primary importance. For example, when $m_i = 1$, $Y_i = 1$ may denote the occurrence of a particular event of interest. Large α's (relative to their standard errors) correspond to factors which increase the chance of the event.

There are several reasons for the widespread popularity of the logistic regression model. Cox (1970) argues from considerations of sufficiency. By writing down the likelihood based on $\{(y_1, x_1), \ldots, (y_n, x_n)\}$, one discovers that the vector

$$(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i x_{i,1}, \ldots, \sum_{i=1}^{n} y_i x_{ir})$$

is sufficient for α. Cox(1970) feels that this model is the most useful analogue, for binomial data, of the normal linear model. When the covariates are nominal or ordinal, there is a correspondence between the logistic parameters and the parameters of a log linear model for cross-classified data (Fienberg(1981)). Finally, inference for the α's remains unaffected regardless of whether the data are sampled prospectively or retrospectively (see for examples McCullagh and Nelder (1989)).

The logistic regression model is an example of a Generalized Linear Model (GLM) which is discussed by McCullagh and Nelder (1989). GLMs are models for regression data, i.e. a response Y measured along with a vector of covariates x. Under the GLM formulation, the response Y has a distribution which is a member of the exponential family and some monotonic differentiable function of the expected value $\mu = E(Y)$ (called the link function), $g(\mu)$, is expressed as a linear combination of covariates and parameters. For binomial regression data, the proportion Y/m is regarded as the response and $E(Y/m) = \pi$. Hence for the logistic regression model, the link function is the logit function, i.e., $g(\pi) = \log(\pi/(1-\pi))$

When the logistic regression model fits the data poorly there are several alternative models to consider. Using the GLM formulation, these alternatives can be dichotomized into link function or frequency distribution modifications. To understand some of these generalizations, it becomes important to distinguish between two types of data sets. Suppose that in a designed experiment, experiment units are sampled and a 0-1 response along with some covariates are recorded for each unit. We call such data sets as ungrouped or point binomial, and the fundamental experimental units as Bernoulli experimental ones. Observations of this nature arise, for instance, in some medical trials where an end-period result for each patient (experimental unit) is either recovered ($Y = 1$) or unrecovered ($Y = 0$). Alternatively, if 0-1 responses are grouped under each experimental condition and the cumulative number of positive responses for each condition are recorded along with a vector of covariates describing the condition, we call such data sets from the experiment as grouped and the fundamental experimental units as binomial ones. In a toxicity experiment, for example, tanks of fish are exposed to some toxic agent at several levels and the incidence of liver tumors in each tank is recorded. Here the tumor rates are the fundamental experimental units and each provides a $0, 1, \ldots, m$ response where m_i is the size of the ith tank. With the logistic regression model, this distinction between these two data sets is superfluous. The log-likelihoods under the two regimes differ only by an irrelevant constant term $\sum_{i=1}^{n} \ln \begin{pmatrix} m_i \\ y_i \end{pmatrix}$, and inference remains unaffected. When considering generalizations, however, the distinction between two types of data can be crucial. While grouped data can be modelled by non-binomial frequency distributions, with ungrouped data we do not have this option. Any model for a Bernoulli response, $Y = 0$ or 1, is determined by $P(Y = 1)$, which specifies a binomial model with $m = 1$ and $\pi = P(Y = 1)$.

3.1.1 Link Modifications

A wide choice of link function $g(\pi)$ is available. In addition to the logistic function, at least two other functions are commonly used in practice: (1) the probit function $g(\pi) = \Phi^{-1}(\pi)$ where $\Phi^{-1}(\pi)$ is the inverse of the standard Normal integral. This function is symmetric in π and for any value of π in the range $(0, 1)$, the corresponding value of the probit of π will lie between $-\infty$ and ∞. Note that when $\pi = 0.5$, $probit(\pi) = 0$; and (2) the complementary or log-log complementary function $\ln(-\ln(1-\pi))$. This function again transforms a probability in the range $(0, 1)$ to a value in $(-\infty, \infty)$, but unlike logistic and probit transformations, this function is not symmetric about $\pi = 0.5$. Note that all the three link functions can be regarded as special cases of a general procedure that relates the probability of a positive response to the covariates through a link $G^{-1}(\pi)$ where G is some continuous distribution function. In fact, the logistic link is the inverse of the logistic distribution which is defined as $\pi(z) = \exp(z)/(1+\exp(z)) = Pr(Z < z)$ where Z is a standard logistic random variable. Similarly, the complementary link can be derived by taking the inverse of the extreme value distribution function as the link function.

McCullagh and Nelder (1989) discuss and compare these link functions. Of these three link functions, the use of the complementary function is limited to those situations where it is appropriate to deal with success probabilities in an asymmetric manner. The logit and probit link functions are quite similar to each other, but from computational viewpoint, the logistic transformation is more convenient because it has an explicitly analytical form. There are two other reasons why the logit link function is preferred to the other two link functions. First, it has a direct interpretation in terms of the logarithm of the odds in favor of a success. Second, models based on the logit link function are particularly appropriate for analysis of data that have been collected retrospectively, such as in a case-control study.

Other links include the angular, $g(\pi) = \sin^{-1}(\pi)^{1/2}$ and the linear, $g(\pi) = \pi$. These links are discussed in Cox (1970). Of the links discussed above, the linear, angular, probit and logit are symmetric in the sense that $g^{-1}(z) = 1 - g^{-1}(-z)$ and these links are similar for probabilities in the range $(0.1, 0.9)$.

Relaxing the requirement that $g(\pi)$ be a linear function of the covariates, we can use nonlinear link functions to obtain a richer class of probability functions than the class specified by a linear link. Prentice (1976) generalizes the logistic link symmetrically to

$$\begin{aligned} \pi(x) &= \int_{-\infty}^{x'\alpha} \frac{\exp(w\gamma_1)(1+\exp(w))^{-(\gamma_1+\gamma_2)}}{B(\gamma_1,\gamma_2)} \\ &\equiv \int_{-\infty}^{x'\alpha} f(w)dw, \end{aligned} \tag{3.1}$$

where $B(a, b)$ is the beta function.

When $\gamma_1 = \gamma_2$, inverting (3.1) yields the logistic link. The parameters γ_1 and γ_2 indicate skewness and heaviness of tails of the density $f(w)$. Other special cases of $f(w)$ are extreme minimum value, extreme maximum value, probit, exponential, reflected exponential, and double exponential. Thus this model can be viewed as specifying a richer class of threshold distributions than the logistic alone.

Other link functions include the power transformations of the logit probability (Aranda-Ordaz (1981) and Guerrero and Johnson (1982)). A problem with these nonlinear link functions is that in some cases it may be difficult to compute the maximum likelihood estimates under the corresponding models. With development of high speed computers, this problem may become less important.

Carroll *et al.* (1984) modify the probit link function by including covariates measured with error in Bernoulli experiments. With normal measurement errors, they discuss procedures to compute estimates for this model. They also demonstrate that the usual estimate of the probability of a positive response can be substantially in error when covariates are measured with non-trivial error. Their modification differs from the previous

link alternatives in that the modified link is derived to accommodate a specific problem.

These approaches try to modify or enrich the basic logistic model by focusing on the relationship between the covariates and the probability of a positive response.

3.1.2 Frequency Distribution Modifications

A consequence of using the binomial frequency distribution in the logistic regression is that $Var(Y) = m\pi(x,\alpha)(1-\pi(x,\alpha))$. In practice, however, we often have $Var(Y) > m\pi(x,\alpha)(1-\pi(x,\alpha))$, suggesting the need for alternative frequency distributions. This may be reflected in over-large residual deviance and adjusted residuals which have a variance > 1. We note that if a positive response Y can be expressed as the sum of m independent Bernoulli random variables each with success probability $(\pi(x,\alpha))$, $Var(Y) = m\pi(x,\alpha)(1-\pi(x,\alpha))$. Hence, to use a non-binomial frequency distribution implicitly requires viewing Y as the fundamental response, that is, to have binomial experimental units. Several researchers have proposed approaches to accommodate extra-binomial variability.

Without covariates, an alternative frequency distribution is the beta-binomial distribution

$$f(y \mid a,b,m) = \int_0^1 \binom{m}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{y+a+1}(1-\pi)^{m-y+b-1} d\pi. \tag{3.2}$$

The model is derived by assuming that the binomial parameter π is a positive random variable following a beta(a,b) mixing distribution. Hence the marginal distribution of the response Y is the beta-binomial. Williams (1975) discusses this model for the data from completely randomized toxicological experiments in which the experimental units are animal litters. In the model, the number of deaths among pups within a litter is assumed to have a beta-binomial distribution. This is a sensible situation to consider

binomial generalizations because litter mates often tend to respond more alike than pups from different litters and a binomial model assumes independence between litter mates.

Several researchers generalize the beta-binomial distribution to incorporate covariates in the parameters for some particular applications. Crowder (1978) generalizes the beta-binomial model for 1 and 2 way layouts. It is not obvious, however, how his approach generalizes to continuous covariates. A difficulty with generalizing the beta-binomial to allow more complicated settings, for example continuous covariates, is that one ought to somehow relate the beta-binomial parameter a and b to a covariate vector x via some functions $a(x)$ and $b(x)$. As Ochi and Prentice (1984) point out, it is hard to specify such functions with intuitive appeal.

Otake and Prentice (1984) model the number of aberrant cells in samples of 100 cells taken from human survivors of the atom bombings of Hiroshima and Nagasaki. Possibly due to measurement error of the radiation doses, the data exhibit extra-binomial variability. At each unique x vector, they estimate $a(x)$, $b(x)$ by maximum likelihood using the beta-binomial model of equation (3.2). They then fit a linear model $\bar{Y}(x) = x'\alpha$ via weighted least squares where $\bar{Y}(x)$ is the average number of responses at covariate vector x. The weights are the inverses of the estimated variance of $\bar{Y}(x)$ (based on $\hat{a}(x)$, $\hat{b}(x)$) under the beta-binomial model. They point out that failure to accommodate this variability results in overly precise inference concerning the α's.

Pierce and Sands (1975) used a different approach. They assume that unmeasured covariates or measurement errors might have an additive random effect on the log-odds scale, and that $logit(\pi_i) = x_i'\alpha$ where the intercept α_0 is distributed as a normal (μ, σ^2) random variable. Likelihood estimation and residual analysis are discussed as well as an approximate analysis necessitated by the complicated nature the likelihood function.

Efron (1986) introduces double exponential families as constituent distributions in GLMs, in which means and variances are allowed to depend on covariates. As an example

of his model, he modifies the binomial distribution (m, π) by rescaling it with sample size m to define a double binomial family

$$f(y \mid \pi, \theta, m) = c(\pi, \theta, m)\theta^{1/2}\{g_{\pi,m}(y)\}^{\theta}\{g_{y,m}(y)\}^{1-\theta}[dG_m(y)],$$

where

$$g_{\pi,m}(y) = \binom{m}{my}\pi^{my}(1-\pi)^{m(1-y)}, \qquad y = 0, 1/m, \ldots, 1,$$

and $G_m(y)$ is the discrete distribution putting mass $\binom{m}{my}2^{-m}$ at $y = 0, 1/m, \ldots, 1$, and $c(\pi, \theta, m)$ satisfies

$$\int_{-\infty}^{\infty} f(y \mid \pi, \theta, m)dG_m(y) = 1.$$

Based this model, he analyzes the toxoplasmosis data by incorporating covariates to the mean and variance in such a way that $logit(\pi_i) = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3$ where x_i is the standardized rainfall for city i, and $\theta_i = 1.25/(1+\exp(-\lambda_i))$ where $\lambda_i = \beta_0 + \beta_1 M_i + \beta_2 M_i^2$ and M_i is the standardized value of the sample size m_i for city i.

Another approach for modifying the binomial frequency distribution is quasi-likelihood which specifies only the first two moments of Y rather than the complete distribution. The attraction is that unduly rigorous assumptions about the frequency distribution are avoided. To model binomial a regression $(y_1, x_1), \ldots, (y_n, x_n)$, McCullagh and Nelder (1983) suggest assuming that $E(Y_i) = m_i\pi_i$ and $Var(Y_i) = m_i\sigma^2\pi_i(1-\pi_i)$ rather than specifying a complete distribution for Y, where $\pi = \pi(x_i, \alpha)$. This approach is similar to one advocated by Finney (1971) who used the probit instead of the logit link. Note that for the logistic regression model, $\sigma^2 = 1$. Therefore $\sigma^2 > 1$ corresponds to extra-binomial variability or overdispersion, while $\sigma^2 < 1$ corresponds to underdispersion. Since the complete distribution of Y is not specified, maximum likelihood estimation is precluded.

Estimates of α and σ^2 are computed via a quasi-likelihood (Wedderburn, 1974) approach. In fact, the maximum quasi-likelihood estimates of α are the same as the usual logistic regression maximum likelihood estimates regardless of the value of σ^2, and the moment estimate of σ^2 equals Pearson's chi-square value divided by the degree of freedom. This estimate is consistent in the limit as the number of observations increases to infinity with m_i fixed, and its asymptotic distribution is known (McCullagh and Nelder (1983)). Another estimate of σ^2 is obtained by the deviance divided by degree of freedom. A problem with this model is lack of interpretation of it because it cannot explain the cause of overdispersion as other quasi-likelihood models such as that of Williams does.

Williams (1982) considers two quasi-likelihood models which fine tune the previous approach. By regarding the binomial parameter π_i as an unspecified random variable Π_i following a continuous mixing distribution on $(0,1)$ with $E(\Pi_i) = \theta_i$ and $Var(\Pi_i) = \phi\theta_i(1-\theta_i)$, he shows that the unconditional mean and variance of Y_i are

$$\begin{aligned} E(Y_i) &= m_i\theta_i \qquad \text{and} \\ Var(Y_i) &= m_i\theta_i(1-\theta_i))(1+\phi(m_i-1)), \end{aligned}$$

where $\theta_i = \exp(x_i'\alpha)/(1+\exp(x_i'\alpha)$. Note that in the absence of random variation in the response probabilities, Y_i would have a binomial distribution, $Bi(m_i, \theta_i)$, and in this case, $Var(Y_i) = m_i\theta_i(1-\theta_i)$. This corresponds to the situation where $\phi = 0$ in the above equation. On the other hand, if there is variation of Y_i amongst the response probabilities, so that ϕ is greater than zero, the unconditional variance of Y_i will exceed $m_i\theta_i(1-\theta_i)$ by a factor $(1+\phi(m_i-1))$. Thus variation amongst the response probabilities causes the variance of the observed number of successes to be greater than it would have been if the response probabilities did not vary at random, resulting in overdispersion.

As Ochi and Prentice (1984) and Collett (1991) mention, this model can be also derived by assuming that there is a common correlation between the Bernoulli responses

within a binomial experimental units. Suppose that the ith of m sets of binary data consists of Y_i successes in m_i observations. Let $R_{i1}, \ldots, R_{im_i}$ be the random variables associated with the m_i observations in this set, where $R_{ij} = 1$ for $j = 1, \ldots, m_i$, corresponds to a success, and $R_{ij} = 0$ to a failure. Now suppose that the probability of a success is θ_i, so that $P(R_{ij} = 1) = \theta_i$, $E(R_{ij}) = \theta_i$ and $Var(R_{ij}) = \theta_i(1 - \theta_i)$. The number of successes Y_i is then the random variable $\sum_{j=1}^{m_i} R_{ij}$, and so $E(Y_i) = \sum_{j=1}^{m_i} E(R_{ij}) = m_i\theta_i$, and the variance of Y_i is given by

$$Var(Y_i) = \sum_{j=1}^{m_i} Var(R_{ij}) + \sum_{j=1}^{m_i} \sum_{k \neq j} Cov(R_{ij}, R_{ik})$$

where $Cov(R_{ij}, R_{ik})$ is the covariance between R_{ij} and R_{ik} for $j \neq k$, and $k = 1, \ldots, m_i$. If the m_i random variables $R_{i1}, \ldots, R_{im_i}$ were mutually independent, each of these covariance terms would be zero. However, since we assume that the correlation between R_{ij} and R_{ik} is

$$\delta = \frac{Cov(R_{ij}, R_{ik})}{\sqrt{Var(R_{ij})Var(R_{ik})}},$$

we have $Cov(R_{ij}, R_{ik}) = \delta\theta_i(1 - \theta_i)$ and

$$\begin{aligned} Var(Y_i) &= \sum_{j=1}^{m_i} \theta_i(1 - \theta_i) + \sum_{j=1}^{m_i} \sum_{k \neq j} \delta\theta_i(1 - \theta_i) \\ &= m_i\theta_i(1 - \theta_i) + m_i(m_i - 1)[\delta\theta_i(1 - \theta_i)] \\ &= m_i\theta_i(1 - \theta_i)[1 + (m_i - 1)\delta]. \end{aligned}$$

Note that the approach of McCullagh and Nelder lacks this interpretation unless $m_i = m$ for $i = 1, \ldots, n$. An iterative algorithm which produces estimates of α and π is also presented. Unlike the approach of McCullagh and Nelder, the estimates of α may be different from the usual logistic regression maximum likelihood estimates unless $m_i = m$ for $i = 1, \ldots, n$.

Williams (1982) also discusses another model where the logit of π_i is a random variable with $E(logit(\pi_i)) = x'\alpha$ and $Var(logit(\pi_i)) = \sigma^2$. As a consequence of this assumption, the true response probability is a random variable Π whose expected value is θ_i. The resulting model for $logit(\Pi_i)$ is then

$$logit(\Pi_i) = x_i\alpha + \delta_i$$

and the term δ_i is known as a random effect. This model generalizes the approach of Pierce and Sands by relaxing the assumption that the intercept of the regression has a normal distribution. Williams(1982) notes that these two models are quite similar though the latter has a more elegant interpretation since the fixed and random effects are on the same scale.

Follmann and Lambert (1989) propose a non-parametric mixture of logistic regression model in which the intercept in the regression is a random variable with an unknown mixing probability distribution, and other regression coefficients are unknown constants. The mixed probability function of the response Y associated with a covariate vector x and m trials is given by

$$f(y \mid x, \alpha, m, H) = \binom{m}{y} \int_{-\infty}^{\infty} \pi(a + x'\alpha)^y (1 - \pi(a + x'\alpha))^{m-y} dH(a), \tag{3.3}$$

where $\pi(a + x'\alpha) = \exp(a + x'\alpha)/(1 + \exp(a + x'\alpha))$ and $y = 0, 1, \ldots, m$.

Although the mixing distribution H is not indexed by parameters, Laird (1978) has shown, under general conditions, that when estimating any mixture model (without covariates), the nonparametric maximum likelihood estimator of H is a step function with a finite number of steps. Lindsay (1983) also discusses some general results for nonparametric mixtures. He shows that existence, uniqueness and support size of the maximum likelihood estimate are related to properties of the convex hull of the likelihood. These results given by Laird (1978) and Lindsay (1983) imply that in terms of maximum

likelihood estimate, it is the same no matter whether H is assumed as a nonparametric distribution or as a discrete distribution with c points of support, where c is an unknown finite integer. In this sense, (3.3) may be equivalently expressed by a finite mixture with an unknown number of components. In the next section we will propose a mixed logistic regression model which generalizes Follmann and Lambert's model.

3.2 Tests For Extra-binomial Variation

To check whether data are overdispersed relative to the binomial assumption, we need a way to test for extra-binomial variation for regression type data. Note that it may be misleading if one tests for extra-binomial variation by fitting a more comprehensive model that includes the binomial, and tests a reduction to the simple model using, for instance, a likelihood ratio test. Lawless (1987a) points out that in some circumstances the asymptotic distribution used with these cases may be unreliable, as they tend to underestimate the evidence against the base model.

An informal approach to detect extra-binomial variation is to use convexity plots (Lindsay and Roeder, 1992, and Lambert and Roeder, 1993). For example, Lambert and Roeder (1993) define the following function $C(\pi)$ and propose plotting it against π for logistic regression

$$C(\pi) = n^{-1} \sum_{i=1}^{n} \left(\frac{\pi}{\hat{y}_i}\right)^{y_i} \left(\frac{1-\pi}{1-\hat{y}_i}\right)^{m_i - y_i},$$

where $\hat{y}_i = \exp(x_i'\hat{\alpha})/(1+\exp(x_i'\hat{\alpha}))$, $\hat{\alpha}$ is the maximum likelihood estimate of regression parameter vector α, and $\pi \in (0,1)$. They prove that if observations are generated by a logistic regression model with random coefficients or random means, $C(\pi)$ is approximately convex for a large sample. Therefore, the more convex $C(\pi)$ appears, the more evidence there is of overdispersion or an omitted variable. Note that this approach cannot distinguish overdispersion from lack-of-fit problem.

Several researchers use score tests for extra-binomial variation by fitting the binomial model as a first step in the model building and testing for overdispersion. Tarone (1976) considers a correlated binomial alternative model, and applies the $C(\alpha)$ procedure of Neyman (1959) to derive the score test statistic for the adequacy of the binomial distribution. Taking a different approach, Efron (1986) derives the score test statistic against beta-binomial alternatives. Dean (1992) develops a unifying theory for the score tests mentioned above and provides three score test statistics for the hypotheses of no overdispersion in the usual logistic regression model against alternatives based on three different forms of extra-binomial variation respectively. These score test statistics are

$$N_a = \frac{\sum_{i=1}^n \{(y_i - m_i\hat{\pi}_i)^2 - m_i\hat{\pi}_i(1-\hat{\pi})\}}{\hat{V}},$$

$$N_b = \frac{\sum_{i=1}^n \{[\hat{\pi}_i(1-\hat{\pi}_i)]^{-1}(y_i - m_i\hat{\pi}_i)^2 + \hat{\pi}_i(y_i - m_i\hat{\pi}_i) - y_i(1-\hat{\pi})\}}{\{2\sum_{i=1}^n m_i(m_i-1)\}^{1/2}} \quad \text{and}$$

$$N_c = \frac{\sum_{i=1}^n \{(m_i-1)\hat{\pi}_i(1-\hat{\pi}_i)\}^{-1}\{(y_i - m_i\hat{\pi}_i)^2 + \hat{\pi}_i(y_i - m_i\hat{\pi}_i) - y_i(1-\hat{\pi})\}}{\{2\sum_{i=1}^n m_i(m_i-1)^{-1}\}^{1/2}}$$

corresponding to the following specifications of overdispersion:

(a) $E(Y_i) \simeq m_i\pi_i$ and $Var(Y_i) \simeq m_i\pi_i(1-\pi)[1+\theta(m_i-1)\pi_i(1-\pi_i)$ for θ small,

(b) $E(Y_i) = m_i\pi_i$ and $Var(Y_i) = m_i\pi_i(1-\pi)[1+\theta(m_i-1)]$, and

(c) $E(Y_i) = m_i\pi_i$ and $Var(Y_i) = m_i\pi_i(1-\pi_i)(1+\theta)$ for $\theta > 0$.

In the formula for N_a, $\hat{V}$ is calculated by

$$V^2 = \frac{1}{4}\sum_{i=1}^n \{2m_i^2\pi_i^2(1-\pi_i)^2 + m_i\pi_i(1-\pi_i)(1-6\pi_i+6\pi_i^2)\} - \frac{1}{4}\sum_{i=1}^n\sum_{j=1}^n h_{ij}\frac{W_{2i}W_{2j}}{\sqrt{W_{1i}W_{1j}}},$$

where $W_{1i} = m_i\pi_i(1-\pi_i)$, $W_{2i} = \frac{1}{2}m_i\pi_i(1-\pi_i)(1-2\pi_i)$, and h_{ij} are the elements of the matrix $H = W_1^{1/2}X(X^tW_1X)^{-1}W_1^{1/2}$ where $W_1 = diag\{W_{11},\ldots,W_{1n}\}$ and X is the design matrix. In the above three formulae $\hat{\pi}_i$ is the estimated probabilities for positive

response for the independent identical observations based on the usual logistic regression. Under the null hypothesis H_0: $\theta = 0$, each statistic asymptotically follows the standard normal distribution. Note that the first two types of overdispersion (a) and (b) are the mean-variance relationship of the models proposed by Williams (1982), and (c) is that introduced by McCullagh and Nelder (1983).

3.3 A Mixed Logistic Regression Model

Without covariates the finite mixture approach has been widely used in many applications (c.f. Titterington et. al. (1985)). With covariates, however, this approach has not been systematically studied and directly applied for analyzing binomial response data. In this section we extend the finite binomial mixture model to the logistic regression model by allowing both the component binomial parameters and mixing probabilities to depend on covariates. We investigate some basic features of the model. We also discuss identifiability for the model and provide sufficient conditions for identifiability.

3.3.1 The Model

Let the random variable Y_i denote the ith binomial response variable, and let $\{(y_i, m_i, x_i), i = 1, \ldots, n\}$ denote observations where y_i are observed value of Y_i, m_i are total trials for Y_i, and $x_i = (x_i^{(m)}, x_i^{(r)})'$ are k-dimensional covariate vectors associated with y_i. Note that $x_i^{(m)}$ and $x_i^{(r)}$ are k_1-dimensional and k_2-dimensional vectors corresponding to the regression part of mixing probabilities and component binomial parameters respectively. Usually the first element of $x_i^{(m)}$ and $x_i^{(r)}$ is 1 corresponding to an intercept. Our mixed logistic regression model assumes

(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 which generates y_i. Further, the (Y_i, Π_i) are pairwisely independent;

(3) Conditional on covariate $x_i^{(m)}$, Π_i follows a discrete distribution with c points of support, and $Pr(\Pi_i = j \mid x_i^{(m)}, \beta) = p_j(x_i^{(m)}, \beta)$ where $\sum_{j=1}^{c} p_j(x_i^{(m)}, \beta) = 1$ and $p_j(x_i^{(m)}, \beta)$ is defined by

$$\begin{aligned} p_j(x_i^{(m)}, \beta) &\equiv p_{ij} \\ &= \frac{\exp(\beta_j' x_i^{(m)})}{1 + \sum_{k=1}^{c-1} \exp(\beta_k' x_i^{(m)})} \quad \text{for } j = 1, \ldots, c-1, \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} p_{ic} &\equiv p_c(x_i^{(m)}, \beta) \\ &= 1 - \sum_{j=1}^{c-1} p_{ij} \end{aligned} \tag{3.5}$$

with $\beta = (\beta_1, \ldots, \beta_{c-1})'$ and $\beta_j = (\beta_{j1}, \ldots, \beta_{jk_1})'$, $j = 1, \ldots, c-1$, are unknown parameters. In fact, conditional on $x_i^{(m)}$, Π_i follows a multinomial distribution $(1, p_{i1}, \ldots, p_{ic})$. Note that β appears in each p_{ij} for $1 \leq j \leq c$;

(4) Conditional on $\Pi_i = j$ and the binomial denominator m_i, Y_i follows a binomial distribution which we denote by

$$\begin{aligned} Y_i &\sim f_j\left(y_i \mid x_i^{(r)}, m_i, \alpha_j\right) \\ &\equiv \mathrm{bi}\,(y_i \mid m_i, \pi_{ij}) \\ &= \binom{m_i}{y_i} \pi_{ij}^{y_i} (1 - \pi_{ij})^{m_i - y_i} \end{aligned} \tag{3.6}$$

where

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

where $\alpha \equiv (\alpha_1, \ldots, \alpha_c)'$ are unknown parameters , where $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jk_2})'$, $j = 1, \ldots, c$. Note that the component binomial parameter π_{ij} relate to covariates $x_i^{(r)}$ through the logit function.

The above assumptions define the unconditional distribution of observations, y_i, as a finite binomial mixture in which the mixing probabilities, p_{ij}, depend on the covariates $x_i^{(m)}$ through the multinomial link function, and the component distributions are binomial distributions with the probabilities, π_{ij}, depending on the covariates $x_i^{(r)}$ through the logit function. Suppose that observations can be classified into c groups corresponding to the c unobservable states, α_j may be interpreted as the coefficients of the logistic regression of observations in group j. On the other hand, β 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 the 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, i.e., $\alpha_{jl} = 0$ for some j, $j = 1, \ldots, c$. We denote $X^{(m)} = (x_1^{(m)} \ldots x_n^{(m)})'$ and $X^{(r)} = (x_1^{(r)} \ldots x^{(r)})'$ as two design matrices.

Under the above assumptions the probability function of Y_i satisfies

$$f(y_i \mid x_i^{(r)}, x_i^{(m)}, m_i, \alpha, \beta) = \sum_{j=1}^{c} p_{ij} \mathrm{bi}\,(y_i \mid m_i, \pi_{ij}) \tag{3.7}$$

where p_{ij} and $\mathrm{bi}\,(y_i \mid m_i, \pi_{ij})$ are specified by (3.4),(3.5) and (3.6) 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 m_i trials, and probability π_{ij}.

There are several justifications for mixed logistic regression models. Suppose that each experiment unit or object has some underlying propensity for a positive response

which is captured by one of the c response curves: $\text{logit}(\pi) = x^{(r)'}\alpha_j$, $(1 \leq j \leq c)$, and that the proportion of the experiment units captured by the jth curve, depends on a covariate vector $x^{(m)}$, i.e., $p_j(x^{(m)}, \beta)$. Thus we are led to the model of equation (3.7).

Another argument for the mixed logistic regression model is that the coefficient vector α in the usual logistic regression model, $\text{logit}(\pi) = x^{(r)'}\alpha$, is a random variable with the discrete distribution: $Pr(\alpha = \alpha_j) = p_j$ for $j = 1, \ldots, c$. By making the further assumption that p_j are related to a covariate vector $x^{(m)}$ we are led to the model of equation (3.7).

Note that the above model includes several interesting models as special cases. Some of them were previously studied.

- Choosing $c = 1$ yields a logistic regression model;
- Setting $x_i^{(m)} = (1)$ yields a finite mixed logistic regression model with constant mixing probabilities;
- Setting $x_i^{(m)} = (1)$, $x_{i1}^{(r)} \equiv 1$ and $\alpha_{jk} \equiv \alpha_k$ for $k \neq 1$ yields Follmann and Lambert's model (1989);
- Setting $x_i^{(r)} = (1)$ yields the finite binomial mixture in which the component binomial parameters are constant and the mixing probabilities depend on covariates $x_i^{(m)}$.

3.3.2 Features of the Mixed Logistic Regression Models

To use the mixed logistic regression models we have to distinguish experiment units as either Bernoulli or binomial. For binary data ($m_i = 1$ for $1 \leq i \leq n$) we can rewrite

equation (3.7) as

$$f(y \mid x_i^{(r)}, x_i^{(m)}, m_i, \alpha, \beta) = [\sum_{j=1}^{c} p_{ij}\pi_{ij}]^y [1 - \sum_{j=1}^{c} p_{ij}\pi_{ij}]^{1-y}.$$

The above equation implies that we only modify the link function with the probability $\pi_i = \sum_{j=1}^{c} p_{ij}\pi_{ij}$. In this case, no matter whether the binary responses are heterogeneous, the responses always have Bernoulli distributions. This also means that the above model cannot adjust for overdispersion relative to the Bernoulli assumption. Furthermore, the model may not be identifiable without imposing some unrealistic restrictions on covariates. Hence we recommend not using the mixed logistic regression models when dealing with binary data.

For binomial experimental units, the distribution defined by equation (3.7) is no longer a member of exponential family so that the representation of a generalized linear model does not apply. In this case the component distributions have the logistic link, and the frequency distribution is a finite binomial mixture.

For the mixed logistic regression models, the unconditional mean and variance of Y_i are, respectively,

$$\begin{aligned} E(Y_i) &= E(E(Y_i \mid \Pi_i)) \\ &= m_i(\sum_{j=1}^{c} p_{ij}\pi_{ij}) \equiv m_i\tilde{\pi}_i \end{aligned} \tag{3.8}$$

and

$$\begin{aligned} Var(Y_i) &= E(Var(Y_i \mid \Pi_i)) + Var(E(Y_i \mid \Pi_i)) \\ &= m_i \sum_{j=1}^{c} p_{ij}\pi_{ij}(1 - \sum_{j=1}^{c} p_{ij}\pi_{ij}) + ((m_i - 1)/m_i)\, Var(E(Y_i \mid \Pi_i)) \\ &= m_i\tilde{\pi}_i(1 - \tilde{\pi}_i) + ((m_i - 1)/m_i)\, Var(E(Y_i \mid \Pi_i)), \end{aligned} \tag{3.9}$$

where

$$Var(E(Y_i \mid \Pi_i)) = m_i^2 \left\{ \sum_{j=1}^{c} p_{ij}\pi_{ij}^2 - (\sum_{j=1}^{c} p_{ij}\pi_{ij})^2 \right\}.$$

Since $m_i > 1$, $Var(E(Y_i \mid \Pi_i)) = 0$ holds if and only if $E(Y_i \mid \Pi_i)$ is constant. Hence, if we denote $\tilde{\pi}_i$ as the new probability, $Var(Y_i) = m_i\tilde{\pi}_i(1-\tilde{\pi}_i)$ if and only if $\pi_{i1} = \ldots = \pi_{ic}$ for $1 \leq i \leq n$. This implies that the proposed model is able to cope with extra-binomial variation among $Y_1, \ldots, Y_n$ due to heterogeneity in the population.

3.3.3 Identifiability

To be able to reliably estimate the parameter of (3.7) we require the mixture be identifiable, that is, two sets of parameters which do not agree after permutation cannot yield the same mixture 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 (1961,1963), Blischke (1964) and Margolin, Kim and Risko (1989) give necessary and sufficient conditions for identifiability of the finite binomial mixtures. These results may be summarized as follows. In the binomial family bi(M, π), $0 < \pi < 1$, for fixed M but varying π, the class of mixtures of at most k members is identifiable if and only if $M \geq 2k-1$. That is, if there are two representations of the same mixture:

$$\sum_{j=1}^{c_1} p_j F_j(y) = \sum_{j=1}^{c_2} \tilde{p}_j \tilde{F}_j(y), \quad y = 0, \ldots, M,$$

with $F_j(y) = \text{bi}(y \mid M, \pi_j)$, $\tilde{F}_j(y) = \text{bi}(y \mid M, \tilde{\pi}_j)$, $0 < p_j < 1$ for $1 \leq j \leq c_1$, $\tilde{p}_j$ for $1 \leq j \leq c_2$, $\sum_{j=1}^{c_1} p_j = \sum_{j=1}^{c_2} \tilde{p}_j = 1$ and $c_1, c_2 \leq k$, then

$$c_1 = c_2, \quad \pi_j = \tilde{\pi}_j, \quad \text{and} \quad p_j = \tilde{p}_j \quad \text{for } j = 1, \ldots, c_1,$$

if and only if $k \leq (M+1)/2$.

With covariates, Follmann and Lambert (1991) discuss the sufficient conditions for the identifiability of the nonparametric logistic regression model with common nonrandom regression coefficients and a random intercept with a finite, unknown mixing distribution.

Note that their model may be equivalently viewed as a special case of our models. They show that for binary response the number of components in the mixture must be bounded by a function of the number of covariate vectors that agree except for one coordinate; and for binomial response the number of components must satisfy the same bound or be bounded by a function of the largest number of trials per response ($\bar{M}$), i.e., $c \leq (\bar{M}+1)/2$.

To discuss sufficient conditions for identifiability in our case, we first define identifiability as follows

Definition: Let $\mathcal{G}_c$ denote the class of probability models $\{f(y_1 \mid x_1^{(r)}, x_1^{(m)}, m_1, \alpha, \beta), \ldots, f(y_n \mid x_n^{(r)}, x_n^{(m)}, m_n, \alpha, \beta)\}$, with $f(y_i \mid x_i^{(r)}, x_i^{(m)}, m_i, \alpha, \beta)$ with at most c components, a restriction that $\pi_{11} < \ldots < \pi_{1c}$, parameter space $\mathcal{C} \times \Pi \times \mathcal{P}$, sample spaces $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$, and fixed total number of trials and covariate vectors $(m_1, (x_1^{(r)}, x_1^{(m)})), \ldots, (m_n, (x_n^{(r)}, x_n^{(m)}))$ where $x_i^{(r)} \in R^{k_1}$ and $x_i^{(m)} \in R^{k_2}$ for $i = 1, \ldots, n$. $\mathcal{G}_c$ is *identifiable* if for $(c, \alpha, \beta), (c^*, \alpha^*, \beta^*) \in \mathcal{C} \times \Pi \times \mathcal{P}$,

$$f(y_i \mid x_i^{(r)}, x_i^{(m)}, t_i, \alpha, \beta) = f(y_i \mid x_i^{(r)}, x_i^{(m)}, t_i, \alpha^*, \beta^*) \tag{3.10}$$

for all $y_i \in \mathcal{Y}_i$, $i = 1, \ldots, n$, implies $(c, \alpha, \beta) = (c^*, \alpha^*, \beta^*)$.

Note that the order restriction in the definition means that two models are equivalent if they agree up to permutations of parameters.

Like the setting without covariates, we give sufficient conditions for identifiability by imposing a restriction on c specified by the minimum number of trials for proper subsets of the observations. We state them below.

Theorem 2. Let $S_\lambda = \{(y_{\lambda_i}, m_{\lambda_i}, x_{\lambda_i}); i = 1, \ldots, t \text{ for some } t\}$ denote such a subset of the observations indexed by $\lambda \in \Lambda$ that the ranks of vectors $\{x_{\lambda_1}^{(m)}, \ldots, x_{\lambda_t}^{(m)}\}$ and $\{x_{\lambda_1}^{(r)}, \ldots, x_{\lambda_t}^{(r)}\}$ equal the ranks of the design matrices $X^{(m)}$ and $X^{(r)}$ respectively, and let $N_\lambda = \min\{m_{\lambda_1}, \ldots, m_{\lambda_t}\}$, and $N_{\lambda_0} = \max_{\lambda \in \Lambda}\{N_\lambda\}$. Then $\mathcal{G}_c$ is identifiable if (1)

$c \leq \frac{1}{2}(N_{\lambda_0}+1)$, and (2) $X^{(r)}$ and $X^{(m)}$ are full rank.

Proof. Without loss of generality, we assume that the subset of the first t observations is S_{λ_0} corresponding to N_{λ_0}. Suppose that (c, α, β) and (c^*, α^*, β^*) satisfy equation (3.10), this then implies that for each i and all $y_i \in \mathcal{Y}_i$, $i = 1, \ldots, t$,

$$\sum_{j=1}^{c} p_{ij} \mathrm{bi}\left(y_i \mid m_i, \pi_{ij}\right) = \sum_{j=1}^{c^*} p_{ij}^* \mathrm{bi}\left(y_i \mid m_i, \pi_{ij}^*\right) \tag{3.11}$$

where $p_{ij} \equiv p_j(x_i^{(m)}, \beta)$ and $\pi_{ij} \equiv \pi_j(x_i^{(r)}, \alpha_j)$ are defined above, and p_{ij}^* and π_{ij}^* are defined analogously. Note that each side of equation (3.11) may be regarded as a finite binomial mixture without covariates. Since $c, c^* \leq N \leq m_i$, Teicher's results (1961, 1963) imply that

$$c = c^*, \quad p_{ij} = p_{ij}^* \quad \text{and} \quad \pi_{ij} = \pi_{ij}^* \tag{3.12}$$

for $i = 1, \ldots, t$ and $j = 1, \ldots, c$. By the definition of the model, we obtain

$$\exp(\beta_j' x_i^{(m)}) = \exp(\beta_j^{*\prime} x_i^{(m)}) \quad \text{for } j = 1, \ldots, c-1, \tag{3.13}$$

$$\mathrm{logit}(\alpha_j' x_i^{(r)}) = \mathrm{logit}(\alpha_j^{*\prime} x_i^{(r)}) \quad \text{for } j = 1, \ldots, c, \tag{3.14}$$

for $i = 1, \ldots, t$. Since the logit function is monotone, from (3.13) and (3.14) we obtain

$$\begin{aligned} (\beta_j - \beta_j^*)' x_i^{(m)} &= 0 \quad \text{for } j = 1, \ldots, c-1 \text{ and } i = 1, \ldots, t, \\ (\alpha_j - \alpha_j^*)' x_i^{(r)} &= 0 \quad \text{for } j = 1, \ldots, c \text{ and } i = 1, \ldots, t, \end{aligned}$$

or

$$(\beta_j - \beta_j^*)' X_t^{(m)} = 0 \quad \text{for } j = 1, \ldots, c-1, \tag{3.15}$$

$$(\alpha_j - \alpha_j^*)' X_t^{(r)} = 0 \quad \text{for } j = 1, \ldots, c, \tag{3.16}$$

where $X_t^{(m)}$ and $X_t^{(r)}$ are the submatrices consisting of the first t rows of $X^{(m)}$ and $X^{(r)}$ respectively. Since the ranks of $X_t^{(m)}$ and $X_t^{(r)}$ equal to the ranks of $X^{(m)}$ and $X^{(r)}$ that are full rank, (3.15) and (3.16) imply that $(\alpha, \beta) = (\alpha^*, \beta^*)$. Thus $\mathcal{G}_c$ is identifiable. □

Note that we can assume that condition (2) holds without loss of generality, since if it does not we can reparameterize the model accordingly. Note also that the sufficient conditions for identifiability depend on partial information of the observations.

The conditions in Theorem 2 mean that if the two design matrices are full rank, the mixed logistic regression models are identifiable up to $[(N_{\lambda_0}+1)/2]$ components. For instance, if $N_{\lambda_0}=4$, the theorem only guarantees that one or two-component mixed logistic regression models are identifiable. Note that the sufficient condition $c \leq \frac{1}{2}(N_{\lambda_0}+1)$ may not be the lowest bound for identifiability.

As a simple illustration of Theorem 2, consider the following data in Table 3.1 on the toxicity of ethylene oxide for grain beetles (Busvine, 1938). Note that Follmann and Lambert (1991) discuss identifiability of their model for this data set. We assume a mixed logistic regression model with both binomial parameters and mixing probabilities depending on dose level x_i and an intercept. Hence the ranks of the design matrices $X^{(m)}$ and $X^{(r)}$ are 2. Since any 2×2 submatrix of either $X^{(m)}$ or $X^{(r)}$ is full rank, there are $45\times 45 = 2025$ elements in the index set Λ. N_λ ranges from 24 to 31, and $N_{\lambda_0}=31$. Therefore, Theorem 2 allows 16 components in the mixed logistic regression model. This sufficient condition is the same as that given by Follmann and Lambert (1991).

For two special cases of our model: constant mixing probabilities ($X^{(m)}=\mathbf{1}$) and constant binomial parameters ($X^{(r)}=\mathbf{1}$), the above sufficient conditions can be stated as follows.

Corollary 1. Let $S_\lambda = \{(y_{\lambda_i}, m_{\lambda_i}, x_{\lambda_i}); i=1,\ldots,k_2\}$ denote such a subset of the observations indexed by $\lambda\in\Lambda$ that the rank of vectors $\{x^{(r)}_{\lambda_1},\ldots,x^{(r)}_{\lambda_{k_2}}\}$ equal the ranks of the design matrices $X^{(r)}$. And let $N_\lambda = \min\{m_{\lambda_1},\ldots,m_{\lambda_{k_2}}\}$, and $N_{\lambda_0}=\max_{\lambda\in\Lambda}\{N_\lambda\}$. Then $\mathcal{G}_c$ is identifiable if (1) $c\leq\frac{1}{2}(N_{\lambda_0}+1)$, and (2) $X^{(r)}$ is full rank.

Corollary 2. Let $S_\lambda = \{(y_{\lambda_i}, m_{\lambda_i}, x_{\lambda_i}); i=1,\ldots,k_1\}$ denote such a subset of the

observations indexed by $\lambda \in \Lambda$ that the rank of vectors $\{x_{\lambda_1}^{(m)}, \ldots, x_{\lambda_{k_1}}^{(m)}\}$ equal the ranks of the design matrices $X^{(m)}$. And let $N_\lambda = \min\{m_{\lambda_1}, \ldots, m_{\lambda_{k_1}}\}$, and $N_{\lambda_0} = \max_{\lambda \in \Lambda}\{N_\lambda\}$. Then $\mathcal{G}_c$ is identifiable if (1) $c \leq \frac{1}{2}(N_{\lambda_0} + 1)$, and (2) $X^{(m)}$ is full rank.

3.4 Parameter Estimation

To obtain the maximum likelihood estimates of the parameters in the proposed model requires using an iterative algorithm. Two widely used algorithms can be applied to this case: (1) the EM algorithm (Dempster, Laird and Rubin, 1977) and (2) quasi-Newton algorithms. In this section we discuss how to apply the EM algorithm and the quasi-Newton algorithm to our model with a known number of components. Note that when implementing the EM algorihtm, we also use a quasi-Newton approach for the M-step. We present results of a Monte Carlo study to investigate the performance of our codes and discuss some implementation issues.

3.4.1 The EM algorithm

For a fixed number of components c we obtain maximum likelihood estimates of the parameters in the above model using both the EM algorithm (Dempster, Laird and Rubin, 1977) and the quasi-Newton approach (Nash, 1991). As is now standard in mixture model estimation, we implement the EM algorithm by treating unobservable component membership of the observations as missing data. We discuss choice of number of components below.

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 generated by the mixed logistic regression model. 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

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

where p_{ij} and $\mathrm{bi}\,(y \mid m_i, \pi_{ij})$ are defined by (3.4),(3.5) and (3.6) respectively.

The EM approach finds the maximum likelihood estimates using an iterative procedure consisting of two steps: E-step and 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 which 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 to 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

$$\begin{aligned} \tilde{z}_{ij} &= E\left(z_{ij} \mid Y, M, X^{(m)}, X^{(r)}, \alpha^{(0)}, \beta^{(0)}\right) \\ &= Pr\left(z_{ij} = 1 \mid Y, X^{(m)}, X^{(r)}, M, \alpha^{(0)}, \beta^{(0)}\right) \\ &= \frac{p_j(x_i^{(m)}, \beta^{(0)})\mathrm{bi}(y_i \mid m_i, \pi_j(x_i^{(r)}, \alpha_j^{(0)})}{\sum_{l=1}^{c} p_l(x_i^{(m)}, \beta^{(0)})\mathrm{bi}(y_i \mid m_i, \pi_l(x_i^{(r)}, \alpha_l^{(0)})}, \quad \text{for } j = 1, \ldots, c, \end{aligned} \tag{3.17}$$

where $p_j(x_i^{(m)}, \beta^{(0)})$ and $\text{bi}(y_i \mid m_i, \pi_j(x_i^{(r)}, \alpha_j^{(0)}))$ are defined by (3.4), (3.5) and (3.6) respectively.

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

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

where

$$\begin{aligned} Q_1(\beta \mid \beta^{(0)}) &= \sum_{i=1}^{n} \sum_{j=1}^{c} \tilde{z}_{ij} \log(p_{ij}) \quad \text{and} \\ Q_2(\alpha \mid \alpha^{(0)}) &= \sum_{i=1}^{n} \sum_{j=1}^{c} \tilde{z}_{ij} \log\left(\text{bi}(y_i \mid m_i, \pi_{ij})\right). \end{aligned}$$

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

$$\frac{\partial Q_1}{\partial \alpha} \Big|_{\hat{\alpha}} = 0 \tag{3.18}$$

$$\frac{\partial Q_2}{\partial \beta} \Big|_{\hat{\beta}} = 0. \tag{3.19}$$

Since closed form solutions of these equations are unavailable, we use a quasi-Newton approach (Nash, 1990) to obtain estimates. 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 tolerance 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 (3.17). To avoid overflow problem in the calculation of $\tilde{z}_{ij}$, we divide both the numerator and denominator in (3.17) by the largest term in the sum in the denominator;

Step 2: (M-step) Find values of $\hat{\alpha}$ and $\hat{\beta}$ to solve (3.18) and (3.19), respectively, using the quasi-Newton algorithm (Nash, 1990);

Step 3: 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, stop.

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

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

(3) $\mid l(\hat{\alpha}, \hat{\beta} \mid Y, M, X^{(m)}, X^{(r)}) - l(\alpha^{(0)}, \beta^{(0)} \mid Y, M, X^{(m)}, X^{(r)}) \mid \geq \epsilon_2$, where $l(\alpha, \beta \mid Y, M, X^{(m)}, X^{(r)})$ is the observed likelihood function.

Dempster, Laird and Rubin (1977) and Wu (1983) discussed the convergence properties of the EM algorithm in a general setting. Since $Q(\alpha, \beta \mid \alpha^{(0)}, \beta^{(0)})$ and its first order partial derivatives are continuous in α, β, $\alpha^{(0)}$ and $\beta^{(0)}$, applying Wu's theorems (1983) lets us conclude that the sequence of the observed likelihood $l(\alpha^{(k)}, \beta^{(k)} \mid Y, M, X^{(m)}, X^{(r)})$ converges to a local maximum or saddle point. Note that the observed likelihood function $l(\alpha, \beta \mid Y, M, X^{(m)}, X^{(r)})$ need not, in general, be globally concave. Thus we need to choose initial values carefully in order to increase the chance that the algorithm converges to the global maximum. Our approach will be discussed below.

Note that the above EM algorithm does not directly yield estimates of the standard errors corresponding to the parameter estimates. On the other hand, when c is known, asymptotic normality of $\sqrt{n}((\hat{\alpha}, \hat{\beta}) - (\alpha, \beta))$ is easily proved under standard regularity conditions (Lehmann, 1983). To approximate standard errors, we may compute $\hat{\sigma}(\hat{\alpha}_{j,l})$ and $\hat{\sigma}(\hat{\beta}_{j,l})$ from the diagonal elements of the inverse of the $(c * k_1 + (c-1) * k_2)$ -

dimensional observed information matrix with c fixed at $\hat{c}$ which is defined as

$$\frac{\partial^2 l(\alpha,\beta \mid Y, X^{(r)}, X^{(m)}, M,)}{\partial(\alpha,\beta)^2} = \begin{pmatrix} \frac{\partial^2 l}{\partial\alpha^2} & \frac{\partial^2 l}{\partial\alpha\partial\beta} \\ \\ \frac{\partial^2 l}{\partial\alpha\partial\beta} & \frac{\partial^2 l}{\partial\beta^2} \end{pmatrix}.$$

Although the EM algorithm is relatively robust for the choice of initial values, it has a lower convergence rate than the quasi- Newton algorithms. To balance the trade-off between these two algorithms, we first use the EM algorithm until either the likelihood value does not increase significantly in terms of a given tolerance $epsilon_2$ or the parameter estimates do not change significantly in terms of a given tolerance $epsilon_1$, and then shift to a quasi-Newton algorithm which maximizes the observed likelihood function. In doing so we can obtain approximate standard error of the estimates as by-product of the quasi-Newton approach. Note that in some cases the approximate standard errors by the quasi-Newton approach may not be accurate. Hence we recommend calculating the information matrix numerically whenever possible. We modify the above Step 3 as follows:

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

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

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

(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.

3.4.2 Starting Values

To run the code of the above algorithm, we need to choose the starting values for the parameters in the model. Note that the EM only ensures, under some regularity conditions

(Wu, 1983), that the estimates converge to the local maximum points of the likelihood function for the observed data. Furthermore, since the likelihood function may not be globally concave, the several starting points needed to find the maximum likelihood estimates, $\hat{\alpha}$ and $\hat{\beta}$. We propose the following approach for choosing the starting values.

We assume that c is known. At the first step of the approach it calculates the ratios, $\{y_1/m_1, \ldots, y_n/m_n\}$, divides the set of the ratios into c groups in terms of its percentiles and fits the observed data into a c-component mixture with constant covariates, $x_i^{(m)} = x_i^{(r)} = (1)$ by choosing initial values based on the percentile information. At the second step, if necessary, it fits the observed data into a mixed logistic regression model containing only one regression term in either the success probabilities or the mixing probabilities in such a way that the initial values of the parameters included in the previous mixture model equal the estimates of the corresponding parameters from the previous fitting model, and initial values of the parameters *not* in the previous fitting model are set to a small value, say, 0.00001. This process is iterated until a complete set of initial values for the mixture model is obtained. The motivation of this ad hoc approach is based on the idea of cluster analysis. At each iteration, we use different criteria to classify the data. First, the data are classified in terms of its percentiles. Then the data are classified in terms of a finite binomial mixture without covariates, and subsequently in terms of mixed logistic regression models. Note that choosing a complete set of initial values for a mixture model step by step in such a way guarantees that the likelihood values will increase in each step. Also our approach produces maximum likelihood estimates for a sequence of nested mixture models while it achieves a complete set of initial values for the mixture model.

We use an example to explain this approach. Suppose that we need to choose initial values to fit a 3-component mixture model with covariates $x_i^{(r)} = (1, s_i)$ and $x_i^{(m)} = (1, t_i)$ where s_i and t_i are real numbers, each with a regression term. First, we find 16.5, 33.0,

49.5, 66.0 and 82.5 percentiles of the observed ratios $\{y_1/m_1, \ldots, y_n/m_n\}$ denoted as q_1–q_5 respectively, and fit the data into a 3-component binomial mixture of constant covariates ($x_i^{(r)} = x_i^{(m)} = (1)$) with the initial values of α_{11}, α_{21} and α_{31} equal to $\text{logit}^{-1}(q_1)$, $\text{logit}^{-1}(q_3)$ and $\text{logit}^{-1}(q_5)$ respectively, and both the initial values of β_{11} and β_{21} equal to 0. Note that under this specification and the logit link function, the initial values of $\pi_j(x_i^{(r)}, \alpha_j)$, $(j = 1, 2, 3)$ are equal to q_1, q_3 and q_5 with the same mixing probabilities 1/3. Second, we fit the data into the 3-component mixed logistic regression model with $x_i^{(r)} = (1, s_i)$ and $x_i^{(m)} = (1)$ by choosing the initial values of α_{12}, α_{22} and α_{32} equal to 0.00001 and the initial values of the other parameters equal to the estimates of the corresponding parameters of the first fitting model. Finally, we choose initial values for the 3-component mixed logistic regression model with $x_i^{(r)} = (1, s_i)$ and $x_i^{(m)} = (1, t_i)$ in such a way that β_{12} and β_{22} are equal to 0.00001 and the other parameters is equal to the estimates of corresponding parameters of the second fitting model.

3.4.3 A Monte Carlo Study

We use Monte Carlo methods to examine the performance of the above algorithm. Particularly, we wished to verify the reliability of our code, determine the precision of estimates and investigate some model selection criteria to be discussed below. We use three 3-component mixture models. For each, we analyzed 101 replicates, each with 100 observations.

Two different approaches for choosing initial values are compared in the study. In one, we use the true parameter values of the model generating the observations as initial values in order to determine performance of the algorithm in the best case. The other uses the true parameter values of α_{11}, α_{21} and α_{31} as initial values, chooses initial values of β_{11} and β_{21} according to the approach described in 3.4.2 section, and fits the samples to a 3-component binomial mixture with constant covariates. Then, following the approach

of section 3.4.2, we choose a complete set of initial values for the parameters of the model generating the samples. These two different approaches of choosing initial values lead to essentially the same estimates. We describe the details below.

Model 1: A model with the success probabilities, π_{ij}, of the component binomial distributions, $\text{bi}(y_i \mid m_i, \pi_{ij})$, depending on one time-dependent covariate, with constant mixing probabilities, where $m_i \equiv 30$. For the logistic regression part,

$$x_i^{(r)} = (1,\ s_i), \tag{3.20}$$

where $s_i = 0.2$ for $i = 1, \ldots, 10$, $d_i = 0.4$ for $i = 11, \ldots, 20$, etc., and

$$\alpha \equiv (\alpha_1,\ \ \alpha_2,\ \ \alpha_3) \tag{3.21}$$

where $\alpha_1' = (-1.2962, -0.4505)$, $\alpha_2' = (-1.3148, 1.0811)$ and $\alpha_3' = (0.6973, 0.7499)$. For the mixing part,

$$\begin{aligned} x_i^{(m)} &= 1 \\ \beta &\equiv (\beta_1,\ \ \beta_2) = (-0.9163, -0.5108). \end{aligned}$$

For the success probabilities can be written with the form

$$\pi_1(x_i^{(r)}, \alpha_1) = \text{logit}(-1.2962 - 0.4505 s_t) \tag{3.22}$$

$$\pi_2(x_i^{(r)}, \alpha_2) = \text{logit}(-1.3148 + 1.0811 s_t) \tag{3.23}$$

$$\pi_3(x_i^{(r)}, \alpha_3) = \text{logit}(0.6973 + 0.7499 s_t), \tag{3.24}$$

and the mixing probabilities

$$\begin{aligned} p_1(x_i^{(m)}, \beta) &\equiv 0.2, \\ p_2(x_i^{(m)}, \beta) &\equiv 0.25 \\ \text{and}\quad p_3(x_i^{(m)}, \beta) &\equiv 0.5. \end{aligned}$$

Note that choosing the parameters as the above makes the component distributions easily distinguished. In this model, p_{i1} decreases from 0.3 to 0.1, p_{i2} increases from 0.3 to 0.7 and p_{i3} increases from 0.7 to 0.9. Thus there are no overlap among them.

Model 2: A model with constant success probabilities, π_{ij}, of the component binomial distributions, $\text{bi}(y_i \mid m_i, \pi_{ij})$, and mixing probabilities depending on one time-dependent covariate, where $m_i \equiv 30$. That is, for the logistic regression part,

$$\begin{aligned} x_i^{(r)} &= 1 \\ \alpha &\equiv (\alpha_1, \ \alpha_2, \ \alpha_3) = (-2.1972, -0.8473, 1.3863) \end{aligned}$$

and for the mixing part,

$$x_i^{(m)} = (1, \ s_i) \tag{3.25}$$

where s_i is defined as above, and

$$\beta \equiv (\beta_1, \ \beta_2) \tag{3.26}$$

where $\beta_1' = (-2.1129, 1.6057)$ and $\beta_2' = (-0.9692, 1.3805)$. The positive probabilities, then, are

$$\begin{aligned} \pi_1(x_i^{(r)}, \alpha_1) &\equiv 0.1 \\ \pi_2(x_i^{(r)}, \alpha_2) &\equiv 0.3 \\ \text{and} \quad \pi_3(x_i^{(r)}, \alpha_3) &\equiv 0.8, \end{aligned}$$

and the mixing probabilities are given by

$$p_1(x_i^{(m)}, \beta) = \frac{\exp(-2.1129 + 1.6057 s_i)}{\exp(-2.1129 + 1.6057 s_i) + \exp(-0.9692 + 1.3805 s_i) + 1} \tag{3.27}$$

$$p_2(x_i^{(m)}, \beta) = \frac{\exp(-0.9692 + 1.3805 s_i)}{\exp(-2.1129 + 1.6057 s_i) + \exp(-0.9692 + 1.3805 s_i) + 1} \quad (3.28)$$

$$p_3(x_i^{(m)}, \beta) = \frac{1}{\exp(-2.1129 + 1.6057 s_i) + \exp(-0.9692 + 1.3805 s_i) + 1}. \quad (3.29)$$

Note that choosing the values of β as the above results in that p_{i1} decreases from 0.2 to 0.1, p_{i2} increases from 0.25 to 0.7 and p_{i3} increases from 0.7 to 0.9. They don't overlap.

Model 3: Both the success probabilities and mixing probabilities depend on the covariate s_i. For the regression part, $x_i^{(r)}$, α and $\pi_j(x_i^{(r)}, \alpha_j)$ are given by (3.20), (3.21), (3.22), (3.23) and (3.24) respectively; For the mixing part, $x_i^{(m)}$, β and $p_j(x_i^{(m)}, \beta)$ are given by (3.25), (3.26), (3.27), (3.28) and (3.29) respectively.

We chose the above parameter values so that the success probabilities and mixing probabilities for each component do not overlap. We would expect that in this case, the algorithm would perform well.

We carried out these simulations, each with 100 replicates. The responses y_i were obtained by first generating a uniform (0,1) random number u_i and then assigning $y_i \sim$ binomial(m_i, π_{i1}) if $u_i \leq p_1(x_i^{(m)}, \beta)$;

$y_i \sim$ binomial(m_i, π_{i2}) if $p_1(x_i^{(m)}, \beta) < u_i \leq p_1(x_i^{(m)}, \beta) + p_2(x_i^{(m)}, \beta)$; and $y_i \sim$ binomial(m_i, π_{i3}) if $u_i > p_1(x_i^{(m)}, \beta) + p_2(x_i^{(m)}, \beta)$. Our implementation of the algorithm used FORTRAN version on a Sun SPARC station 1^+.

The results of the Monte Carlo study are presented in Table 3.2 , Table 3.3 and Table 3.4. These tables show that the mean of estimates are very close to the true parameter values in the models, suggesting that the global maximum of the observed likelihood is reached. For model 1, the sample means are quite close to the true values and the standard deviations are relatively small. Although the coefficients of the logistic regression of model 2 are estimated accurately, estimates of mixing probabilities are more variable. This suggests that estimating mixing probability parameters in this model is intrinsically

more difficult than estimating the success probabilities. This agrees with observations in the literature (Titterington et al., 1985; McLachlan and Basford, 1988). Estimates of the parameters of model 3 illustrate the same pattern as in Model 2 where estimates of the mixing probability parameters are more variable than those of success probabilities parameters. Note, however, that although the estimates of mixing probability parameters, $\hat{\beta}$, vary somewhat, the estimated mixing probabilities, $p_j(x_i^{(m)}, \hat{\beta})$, are more precise due to the multimonial link function between the parameters and mixing probabilities.

The average number of the iterations of the EM algorithm for Model 1 is 8.24, 12.35 for Model 2 and 20.2 for Model 3 under the stopping criterion $\epsilon = 0.01$, and average time is 12.5, 19.4 and 120.5 seconds respectively.

3.5 Implementation Issues

3.5.1 Model Selection

We need to address following the three issues when we apply a mixed logistic regression model: (a) We need to determine the conditions of identifiability for the model; (b) we need to determine the number of components, c, of a mixture, and (c) we need to have a method to carry out inference about model parameters. When c is known, inference for the parameters can be based on a standard likelihood ratio test. In practice, however, this case may not be common. When c is unknown, the usual likelihood ratio test is no longer valid for determining c or testing hypotheses about parameter values. As we discuss in section 2.7.1, this is because mixing probabilities may lie on the boundary of the parameter space when the hypothesized number of components is less than the fitted number of components. Hence the usual regularity conditions for the likelihood ratio test do not hold. We propose the following methods for model selection.

Two widely used model selection criteria are the Akaike's Information Criterion (AIC)

(Akaike, 1973; 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) (see section 2.7.1. For the mixed logistic regression models, we define the AIC and BIC criteria as follows:

- AIC: choose the model for which $\hat{l}_c(X) - a_c(X)$ is largest;
- BIC: choose the model for which $\hat{l}_c(X) - \frac{1}{2}(\log(n))a_c(X)$ is largest

where $\hat{l}_c(X)$ is the maximum log-likelihood of the mixture with c components and covariate X, $a_c(X) = c * k_1 + (c-1) * k_2$ where k_1 and k_2 are the dimensions of α_j and β_j respectively, and n is the total number of observations. These two criteria do not always select the same model.

Using the BIC (AIC), our model selection procedure consists of two stages. At the first stage, we determine c to maximize BIC (AIC) values for the saturated 1-3 (1-4) component mixture models that contain all possible covariates in both success probabilities and mixing probabilities. Note that the c values must be within the range satisfying the identifiability conditions. Although we compute both AIC and BIC values in our applications, we recommend using BIC because our Monte Carlo studies suggest that BIC is more reliable in the model selection. At the second stage, our model selection approach depends on our analysis objectives. If our goal is inference about some particular model parameter, we carry out likelihood ratio tests for nested c-component mixture models. If the goal is choosing an appropriate model to fit the data, we select a model to maximize BIC (AIC) values among c-component mixture model concerned. Since this selection method is heuristic and only gives a guideline in applications, some other specific concerns in model selection should be taken into account from case to case. For instance, in some applications the number of components and some parameters in a mixture may be explicitly or implicitly determined by underlying theory.

In the Monte Carlo studies discussed in Section 3.4.3, we computed both AIC and BIC values for all possible mixed 2 to 4 component models. Table 2.5.1.1 shows that AIC and BIC are reliable methods for choosing the correct models. AIC chose the correct model 94% of the time for Model 1, 82% of the time for Model 2 and 93% of the time for Model 3. When AIC failed to select the correct model, it always chose a model with too many components, suggesting that AIC may under-penalize the number of parameters in the mixtures. On the other hand, BIC *always* chose the correct models, suggesting that BIC may not over-penalize the number of parameters. Note that all sample sizes in the Monte Carlo studies are 100. The examples in the next section will exhibit this procedure in practice.

3.5.2 Classification

One possible use of the mixed logistic regression model is to classify data on the basis of a probabilistic model rather than an ad hoc clustering technique. Since $\tilde{z}_{ij}$ in (3.17) is the estimated posterior probability that the ith observation y_i is generated by the jth component distribution $\mathrm{bi}\,(y_i \mid m_i\pi_{ij})$, this information can be used to classify observations into different groups characterized by the component distributions. For instance, for a c-component mixture model we may postulate c different groups defined by the c different sets of the coefficients of the logistic regression,$\pi_j\left(x_i^{(r)}, \alpha_j\right)$ $(j = 1, \ldots, c)$ of the model. According to the classification criterion, an observation i is identified with the component which maximizes $\tilde{z}_{ij}$. In our applications, maximum values for this quantity all exceed 0.5. Note that if the parameters of the model were known, this classification criterion would be the optimal or Bayes rule (Anderson, 1984, chapter 6) which minimizes the overall error rate. Also such an approach has been referred to as latent class analysis (Aitkin et al. 1981). We illustrate this approach in examples below.

3.5.3 Residual Analysis and Goodness-of-fit

Once a mixed logistic regression model has been fit to a set of observations, it is essential to check the quality of the fit. For this purpose, we first define Pearson, deviance and likelihood residuals for mixed logistic regression models, and then use them to identify individually poorly fitting observations and influential observations on overall fit of the model as well. We also define a quantity to measure influence of individual observations on the set of parameter estimates, and use it to identify influential observations. In addition, we provide goodness-of-fit statistics for mixed logistic regression models.

Definitions of Residuals

As we discuss in Section 2.7.3, we define Pearson, deviance and likelihood residuals for a mixed logistic regression model. The *Pearson residual* is defined as

$$r_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}, \tag{3.30}$$

where

$$\begin{aligned} \hat{\mu}_i &= m_i \sum_{j=1}^{c} \hat{p}_{ij}\hat{\pi}_{ij}, \\ \hat{\pi}_{ij} &= \exp(\hat{\alpha}_j' x_i^{(r)})/(1+\exp(\hat{\alpha}_j' x_i^{(r)})), \\ \hat{p}_{ij} &= \frac{\exp(\hat{\beta}_j' x_i^{(m)})}{\sum_k^{c-1}\exp(\hat{\beta}_k' x_i^{(m)})+1} \quad \text{for } j=1,\ldots,c-1 \text{ and} \\ \hat{p}_{ic} &= \frac{1}{\sum_k^{c-1}\exp(\hat{\beta}_k' x_i^{(m)})+1}, \end{aligned} \tag{3.31}$$

and

$$V(\hat{\mu}_i) = m_i \sum_{j=1}^{c} \hat{p}_{ij}\hat{\pi}_{ij}(1-\sum_{j=1}^{c}\hat{p}_{ij}\hat{\pi}_{ij}) + m_i(m_i-1)\left\{\sum_{j=1}^{c}\hat{p}_{ij}\hat{\pi}_{ij}^2 - [\sum_{j=1}^{c}\hat{p}_{ij}\hat{\pi}_{ij}]^2\right\}. \tag{3.32}$$

The deviance residual is defined as

$$r_{Di} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2[l(y_i,y_i) - l(\hat{\mu}_i,y_i)]}$$

$$= \quad \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \tag{3.33}$$

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

$$D = 2\sum_{i=1}^{n}[l(y_i, y_i) - l(\hat{\mu}_i, y_i)]. \tag{3.34}$$

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

$$\begin{aligned} f(y_i \mid x_i^{(r)}, x_i^{(m)}, m_i, \alpha, \beta) &= \sum_{j=1}^{c} p_{ij}\text{bi}\,(y_i \mid m_i, \pi_{ij}) && (3.35)\\ &\leq \sum_{j=1}^{c} p_{ij}\text{bi}\,(y_i \mid m_i, y_i) && (3.36)\\ &= \text{bi}\,(y_i \mid m_i, y_i) && (3.37) \end{aligned}$$

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

The 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 data set. The value of the likelihood residual for the ith case is defined as

$$r_{Li} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{D - D_{(i)}} \tag{3.38}$$

where $\hat{\mu}_i$ is defined by (3.31); $\hat{\alpha}$ and $\hat{\alpha}_{(i)}$ are the maximum estimates of the regression parameters based on the complete data set of n cases and the data set of $n-1$ cases

excluding the i case respectively; 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 types of residuals approximately follow the standard normal distribution if the fitted model is adequate. Our numerical results show that the Pearson residuals may not be as approximately normal as the other two types of residuals.

Detection of Outliers and Influential Observations

The residuals obtained after fitting a mixed logistic regression model to an observed set of data form the basis of a large number of diagnostic techniques for assessing model adequacy. Since our primary objective of residual analysis for mixed logistic regression models is to identify outliers and influential cases, we discuss how the residuals can be used for this objective.

Like mixed Poisson regression models, we define outliers as those observations that are surprisingly distant from the remaining observations in the sample. Such

observations may occur as a result of measurement errors, that is errors in reading, calculating or recording a numerical value; or they may be just an extreme manifestation of natural variability. Since large residuals indicate poorly fitting observations, we use index plots of residuals for detection of outliers, that is, observations that have unusually large residuals.

The influence of a particular observation on the overall fit of a model can be assessed from the change in the value of a summary measure of goodness of fit that results from excluding the observation from the data set. Since r_{Li}^2 is the change in deviance on omitting the ith observation from the fit, an index plot of these values is the best way of assessing the contribution of each observation to the overall goodness of fit of the model.

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

the following quantity

$$\begin{aligned} w_i &= \frac{1}{p}\left\{\| (\hat{\alpha} - \hat{\alpha}^{(i)})/se(\hat{\alpha}) \| + \| (\hat{\beta} - \hat{\beta}^{(i)})/se(\hat{\beta}) \|\right\} \\ &= \frac{1}{p}\left\{\sum_{j=1}^{c}\sum_{l=1}^{k_1}\frac{|\hat{\alpha}_{j,l} - \hat{\alpha}_{j,l}^{(i)}|}{se(\hat{\alpha}_{j,l})} + \sum_{j=1}^{c-1}\sum_{l=1}^{k_2}\frac{|\hat{\beta}_{j,l} - \hat{\beta}_{j,l}^{(i)}|}{se(\hat{\beta}_{j,l})}\right\} \end{aligned} \tag{3.39}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood parameter estimates of the mixed logistic regression model based on the complete data set of n cases, and $\hat{\alpha}^{(i)}$ and $\hat{\beta}_{(i)}$ on the data set of $n-1$ cases excluding the i case; $se(\hat{\alpha})$ and $se(\hat{\beta})$ are the estimated standard errors of the corresponding estimates based on the n cases, and $p = ck_1 + (c-1)k_2$. Because each term in (3.39) measures a relative change in individual coefficient, w_i can be interpreted average relative coefficient changes 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. The example in the next section will illustrate these points.

Goodness-of-fit Statistics

After fitting a mixed logistic regression model to a set of data, it is natural to enquire about the extent to which the fitted values of the response variable under the model compare with the observed values. If the agreement between the observations and the corresponding fitted values is good, the model may be acceptable. If not the current form of the model will certainly not be acceptable and the model will need to revised. The aspect of the adequacy of a model is widely referred to as goodness of fit.

There are at least two widely used goodness-of-fit statistics which can be used here.

One is the deviance defined as

$$D = \sum_{i=1}^{n} r_{Di}^2 \tag{3.40}$$

where r_{Di} is the deviance residuals for the mixed logistic regression model; and the other is the Pearson's X^2 statistic defined as

$$X^2 = \sum_{i=1}^{n} r_{Pi}^2 \tag{3.41}$$

where r_{Pi} is the Pearson residuals for the mixed logistic regression model. In order to evaluate the extent to which an adopted mixed logistic regression model fits a set of data, the distribution of either the deviance or the Pearson statistic, under the assumption that the model is correct, is needed. In general, the deviance and the Pearson's X^2 statistics are asymptotically distributed as χ^2 with $(n - p)$ degrees of freedom, where n is the number of observations and p is the number of unknown parameters in the model. Many studies have shown that the Pearson statistic is often much more nearly chi-squared than that of the deviance (e.g., Larntz, 1978). For this reason we use the Pearson statistic for overall goodness of fit tests for the mixed logistic regression models.

3.6 An Application

This example uses data from an experiment reported by Ganio and Schafer (1992), which investigates the carcinogenic effects of aflatoxin, a toxic by-product produced by a mold that infects cottonseed meal, peanuts, and grains. Forty tanks of rainbow trout embryos were exposed to either aflatoxin B1 or a related compound, aflatoxicol, at one of five doses for one hour, and the incidence of liver tumors in each tank was recorded after one year. The data in Table 3.5 are the proportions of fish with liver tumors in each of 40 tanks. The researchers believe that there may exist extra-binomial variation due to tank effects and different treatments. They believe that aflatoxical must undergo more

chemical changes than aflatoxin B1 to produce tumors in fish. This may result in more variation of effective doses reaching the liver of fish in aflatoxicol tanks and, therefore, a greater degree of extra-binomial variation for the aflatoxicol group. The issue of interest is to assess dose level and treatment effects on the proportions of fish with liver tumors while taking extra-binomial variation into account.

We first apply the usual logistic regression model with covariates including an intercept, dose level (x_{i1}), treatment (x_{i2}) and dose-treatment interaction (x_{i3}), where

$$x_{i2} = \begin{cases} 0 & \text{if fish in tank } i \text{ was exposed to aflatoxin B1} \\ 1 & \text{if fish in tank } i \text{ was exposed to aflatoxicol} \end{cases} \tag{3.42}$$

$$\text{and} \quad x_{i3} = x_{i1}x_{i2}. \tag{3.43}$$

The top part of Table 3.6 reports results of fitting the data to the usual logistic regression models. Note that the deviance and Pearson goodness-of-fit statistics for the model with covariates x_{i1}, x_{i2} and x_{i3} are 391.08 and 365.3, respectively, with 36 degrees of freedom, suggesting that there is significant evidence of lack of fit in the logistic regression model. Furthermore, the data are overdispersed with respect to the binomial distribution, since each of overdispersion tests is highly significantly ($N_a = 68.26$, $N_b = 36.42$ and $N_c = 36.3$). This also indicates inadequacy of the usual logistic regression model.

Ganio and Schafer (1992) only present exploratory techniques for use in an early stage of data analysis to aid modelling extra-binomial variation. They take some function of the dispersion parameter in a generalized linear model to depend on explanatory variables. To detect extra-binomial variation for the fish data, they consider three models for dispersion. Let π_{ij} be the probability of tumor for concentration level i and carcinogen group j ($i = 1, \ldots, 5$; $j = 1, 2$), and let Y_{ijk} be the number of tumors observed in m_{ijk} fish in tank k of treatment ij. Then they model the variance of this count as $m_{ijk}\pi_{ij}(1-\pi_{ij})/\phi_{ijk}$ and consider the following forms for dispersion parameter: (a) $\phi_{ijk} =$

ϕ; (b) $\phi_{ijk} = \lambda + \alpha z_j$, where $z_j = (j-1)$; and (c) $\phi_{ijk} = [\lambda + \alpha z_{ijk}]^{-1}$, where $z_{ijk} = (m_{ijk} - 1)\pi_{ij}(1 - \pi_{ij})$. Note that Model (a) is a generalized linear model with constant dispersion. Model (b) contains separate dispersion parameters for the two carcinogen groups. Model (c) is the approximate variance of Y if a random effect, with mean 0 and variance α, is additive on the logit scale (Williams, 1982). They find that the extra-binomial variation is associated with the type of carcinogen and cannot be explained simply by differences in the π_{ijk}'s. Note that, however, they do not analyze extra-binomial variation along with dose-response function in mean simultaneously.

We apply the mixed logistic regression model assuming that

(1) each observed number of tumors, y_i, in m_i fish in tank i is associated with covariates $x_i^{(m)} = (1, x_{i1}, x_{i2})$ and $x_i^{(r)} = (1, x_{i1}, x_{i2}, x_{i3})$ where x_{i1}, x_{i2} and x_{i3} are defined above;

(2) numbers of tumors in different tanks are independent and follow a mixed logistic regression model with binomial parameters π_{ij} given by the link function

$$\pi_{ij} = \frac{\exp(\alpha_{j0} + \alpha_{j1}x_{i1} + \alpha_{j2}x_{i2} + \alpha_{j3}x_{i3})}{1 + \exp(\alpha_{j0} + \alpha_{j1}x_{i1} + \alpha_{j2}x_{i2} + \alpha_{j3}x_{i3})}, \tag{3.44}$$

where $i = 1, \ldots, 40$, and $j = 1, \ldots, c$, and the mixing probabilities p_{ij} given by

$$p_{ij} = \frac{\exp(\beta_{j0} + \beta_{j1}x_{i1} + \beta_{j2}x_{i2})}{1 + \sum_{k=1}^{c-1} \exp(\beta_{k0} + \beta_{k1}x_{i1} + \beta_{k2}x_{i2})} \quad \text{for } j = 1, \ldots, c-1, \tag{3.45}$$

and

$$p_{ic} = 1 - \sum_{j=1}^{c-1} p_{ij}. \tag{3.46}$$

Note that since the smallest binomial denominator in the data set is 80, the mixed logistic regression model is identifiable if $c < (80+1)/2 = 40.5$. Thus, there are virtually no restrictions on identifiability in this example.

Table 3.6 provides the results of fitting these models. In order to determine the number of components first, we compare the values of AIC and BIC among the three

staturated models. Clearly, both AIC and BIC lead to the choice of a 2-component mixed logistic regression model. Within these 2-component models we carry out inference using likelihood ratio tests. First we test which covariates in the mixing probabilities are significant. Comparing the model only excluding x_{i2} in mixing probabilities and including all covariates in binomial parameters with the saturated 2-component model, the chi-square test statistic is 0 up to 2 decimal approximation. This clearly indicates that x_{i2} is insignificant in mixing probabilities. Then we test the hypothesis that x_{i1} is insignificant in mixing probabilities. The corresponding chi- square test statistic is $2(1784.36 - 1757.48) = 53.76$ with one degree of freedom, suggesting that x_{i1} is highly significant in mixing probablities.

For binomial parameters we first test the hypothesis that the dose-treatment interaction is insignificant. Comparing the one only excluding x_{i3} in binomial parameters and including x_{i1} in mixing probabilities with the one including all covariates in binomial parameters and x_{i1} in mixing probabilities, the chi- square test statistic is $2(1760.57 - 1757.48) = 6.18$ with 2 degrees of freedom. Since the p-value of the test statistic is 0.0455, we do not reject the hypothesis, at 1% level, that the interaction effect is insignificant. On the other hand, both the effects of covariates x_{i1} and x_{i2} are significant. For instance, to test the hypothesis that the effect of x_{i2} is insignificant, we compare the model including x_{i2} in binomial parameters only and x_{i1} in both mixing probabilities and binomial parameters with the one only including x_{i1} in both mixing probabilities and binomial parameters, and obtain the corresponding chi-square test statistic $2(1834.94 - 1760.57) = 148.74$ with 2 degrees of freedom. Clearly we reject the hypothesis that the effect of x_{i2} is insignificant. Finally, we test the hypothesis of a common effect of treatment for both components, i.e., $\alpha_{12} = \alpha_{22}$. Indeed this hypothesis is valid because the test statistic is 0 up to two decimal approximation. Therefore we select the 2-component mixed logistic regression model with the covariate of dose level

in both mixing probabilities and binomial parameters and the common coefficient of the covariate of treatment in binomial parameters. This model fits the data best.

After fitting the 2-component mixed logistic regression model, the Pearson goodness-of-fit test statistic X^2 is 52.18 with 33 degrees of freedom. The p-value of the test statistic is 0.0181, suggesting that there is no evidence of lack of fit at 1% significance level. Note that the deviance for the fitted model is 51.46 with 33 degrees of freedom. In addition, the Pearson, deviance and likelihood residuals from the fitted model are calculated and displayed in Figure 3.1, Figure 3.2 and Figure 3.3 respectively. These plots show that the three types of residuals are very similar to each other, and that the 37th observation is far distant from the remaining observations in these plots, suggesting that it is an outlier. On omitting the observation, the deviance reduction is $r_{L37}^2 = (-3.1651)^2 = 10.0179$. This means that the 37th observation has great impact on the overall fit of mixed logistic regression model to the data.

For detection of influential observations, the average relative coefficient changes w_i are calculated and displayed in Figure 3.4. Clearly, the 37th observation also has the largest value (0.3543). On omitting this observation, the average relative coefficient change for each parameter estimate is about 35%, and the new parameter estimates become

$$\hat{\beta}_1 = (-44.38, -1183.7) \tag{3.47}$$

$$\hat{\alpha}_1 = (-0.8838, 7.2151, 1.2232) \tag{3.48}$$

$$\hat{\alpha}_2 = (-4.8242, 123.29, 1.2232) \tag{3.49}$$

Note that changes in the binomial parameter estimates for first component are relatively large, while there are almost no changes in the parameter estimates for mixing probabilities. This indicates that the 37th observation has greater influence on the first component than on the second component. We now interpret the fitted model.

The chosen mixed logistic regression model suggests that numbers of fish with liver

tumors are generated by two underlying binomial distributions with binomial parameters defined by, respectively,

$$\pi_{i1} = \frac{\exp(-0.8161 + 6.6209x_{i1} + 1.1686x_{i2})}{1 + \exp(-0.8161 + 6.6209x_{i1} + 1.1686x_{i2})} \tag{3.50}$$

and

$$\pi_{i2} = \frac{\exp(-4.7798 + 122.92x_{i1} + 1.1686x_{i2})}{1 + \exp(-4.7798 + 122.92x_{i1} + 1.1686x_{i2})}. \tag{3.51}$$

In addition, these two distributions are mixed according to the mixing probabilities defined by

$$p_{i1} = \frac{\exp(-44.38 + 1183.7x_{i1})}{1 + \exp(-44.38 + 1183.7x_{i1})} \tag{3.52}$$

and

$$p_{i2} = \frac{1}{1 + \exp(-44.38 + 1183.7x_{i1})}. \tag{3.53}$$

According to this model, tanks in either of the two treatments may be classified into two groups on the basis of the two dose-response functions. For either of the two treatments, fish in those tanks exposed to a higher dose level (> 0.025 pm) follows one dose-response function; and fish exposed to a lower dose level (≤ 0.025 ppm) follows another. In addition, the treatment effect is the same for both groups. On the other hand, when exposed to a higher dose level, there is a higher chance for fish to follow the first dose-response function because the mixing probability for component one is very close to 1. Similarly, when exposed to a lower dose level, there is a higher chance for fish to follow the second dose-response function because the mixing probability for component two is close to 1. Figure 3.5 provides the estimated proportions of fish with liver tumors corresponding to each group for either of the two treatments (the solid line is the proportion for group one and the dotted line for group two). Note that Figure 3.5 also

classifies the observed proportions in terms of the estimated posterior probabilities from the fitted model. Those observations marked as "1" form group one which characterized by the function π_{i1}, while those marked as "2" form group two which is characterized by the function π_{i2}.

Figure 3.6 depicts the mean-variance relationship for the fitted model based on the estimated mean and variance obtained through (3.8) and (3.9). Note that there is no obvious parametric relationship between the estimated mean and variance.

For the purpose of comparison, we also fit the data to the two quasi-likelihood models which are discussed by McCullagh and Nelder (1989) and Williams (1982) respectively. The first assumes a variance form $Var(Y_i) = \sigma^2 m_i \pi_i (1 - \pi_i)$, and the second $Var(Y_i) = m_i \pi_i (1 - \pi_i)[1 + (m_i - 1)\phi]$. Note that the unknown parameters σ^2 and ϕ are usually called unexplained variance. The results of parameter estimates and standard errors are given in Table 3.7. Note that the dose-treatment effect is not significant in quasi-likelihood models (estimates not reported here). As expected, the parameter estimates for both quasi-likelihood models are very similar to each other because the binomial denominators m_i do not vary much. From Table 3.7, we find that parameters estimates under quasi-likelihood models and mixed logistic regression model are different, suggesting that using different methods to model extra-binomial variation may lead to either different parameter estimate or different standard errors or both. For instance, the coefficient estimate for dose level is 12.82 and 12.81 by quasi-likelihood method I and II respectively, and 6.6209 for component one and 122.92 for component two respectively. Furthermore, computing the t-statistic (estimated coefficient/standard error) and comparing the mixed logistic regression model with the quasi-likelihood, we find that quasi-likelihood models may underestimate the treatment and dose effects. For example, the values of the t-statistic of the estimated coefficient for x_{i2} are 4.1556 and 4.1441 for the quasi-likelihood model I and II respectively, while the value for the mixed logistic

regression model is 13.732. Thus, compared with quasi-likelihood methods, the mixed logistic regression model has smaller confidence intervals for parameter estimates.

In summary, we have applied the mixed logistic regression model to analyze the data from a fish toxicology study. The data are well fitted by a 2-component mixed logistic regression model with mixing probabilities depending on the dose level covariate and binomial parameters depending on both dose level and treatment covariates. The goodness-of-fit test suggests that there is no evidence of lack of fit in the model. In addition, the residual analysis identifies an outlier and influential observations. According to this model, there are two dose-response functions for each treatment, which describe lower dose level and higher dose level situations respectively. Comparing with the quasi-likelihood methods, the mixed logistic regression model gives smaller confidence intervals of parameter estimates. Note that both parameter estimates and standard errors under the mixed logistic regression differ from those obtained by the quasi- likelihood method.

3.7 Tables and Figures in Chapter3

Table 3.1: Data of Busvine (1938)

Jar label	Dose	Jar total	Number dead
1	0.033	24	0
2	0.167	31	10
3	0.199	30	17
4	0.225	31	12
5	0.260	27	7
6	0.314	26	23
7	0.322	30	22
8	0.362	31	29
9	0.391	30	30
10	0.394	30	23

Table 3.2: The results of the simulations for the mixed logistic regression model (Model 1)

Parameter	True value	Upper extreme	Upper quartile	Median	Lower quartile	Lower extreme	Average
Initial values set as the true values							
β_{11}	-0.9163	-0.2618	-0.7825	-0.9698	-1.1354	-1.6169	-0.9643
β_{21}	-0.5108	0.0366	-0.3293	-0.5443	-0.7185	-1.2269	-0.5380
α_{11}	-1.2962	-0.6248	-1.0144	-1.2717	-1.4430	-1.9056	-1.2402
α_{12}	-0.4505	0.1955	-0.2766	-0.4593	-0.6294	-0.9752	-0.4855
α_{21}	-1.3148	-0.8427	-1.2082	-1.3796	-1.5097	-1.7884	-1.3619
α_{22}	1.0811	1.4963	1.2325	1.1158	1.0344	0.7893	1.1211
α_{31}	0.6973	1.0054	0.8004	0.6911	0.5779	0.3414	0.6892
α_{32}	0.7499	1.1164	0.8855	0.7695	0.6527	0.4017	0.7683
Initial values chosen step by step							
β_{11}	-0.9163	-0.2467	-0.7711	-0.9626	-1.1331	-1.6162	-0.9586
β_{21}	-0.5108	0.0336	-0.3317	-0.5500	-0.7468	-1.2571	-0.5386
α_{11}	-1.2962	-0.6238	-0.9970	-1.2561	-1.4414	-1.9909	-1.2324
α_{12}	-0.4505	0.1540	-0.3189	-0.4839	-0.6473	-0.9759	-0.4976
α_{21}	-1.3148	-0.7701	-1.1968	-1.3648	-1.5076	-1.7880	-1.2844
α_{22}	1.0811	1.5773	1.2293	1.1066	0.9885	0.6696	1.0599
α_{31}	0.6973	1.0076	0.8131	0.6996	0.5814	0.3416	0.7011
α_{32}	0.7499	1.1165	0.8772	0.7656	0.6473	0.3954	0.7577

Table 3.3: The results of the simulations for the mixed logistic regression model (Model 2).

Parameter	True value	Upper extreme	Upper quartile	Median	Lower quartile	Lower extreme	Average
Initial values set as the true values							
β_{11}	-2.1129	-0.7531	-1.6871	-2.2560	-2.8070	-3.8851	-2.3068
β_{12}	1.6057	2.8955	2.0599	1.7048	1.4069	0.5755	1.7467
β_{21}	-0.9692	0.2170	-0.7619	-1.0517	-1.4186	-2.3082	-1.0958
β_{22}	1.3805	2.3960	1.7416	1.4624	1.1670	0.4227	1.4941
α_{11}	-2.1972	-1.8090	-2.0612	-2.1586	-2.2975	-2.6451	-2.2061
α_{21}	-0.8473	-0.6443	-0.7807	-0.8517	-0.9046	-1.0277	-0.8473
α_{31}	1.3863	1.5637	1.4346	1.3926	1.3371	1.2438	1.3892
Initial values chosen step by step							
β_{11}	-2.1129	-0.7410	-1.6260	-2.2150	-2.7535	-3.8757	-2.2733
β_{12}	1.6057	2.8799	2.0629	1.6947	1.3974	0.5713	1.7469
β_{21}	-0.9692	0.2106	-0.7705	-1.0538	-1.4351	-2.3073	-1.1090
β_{22}	1.3805	2.3804	1.7385	1.4463	1.1490	0.4193	1.4889
α_{11}	-2.1972	-1.6279	-2.0135	-2.1347	-2.2806	-2.6425	-2.1818
α_{21}	-0.8473	-0.6152	-0.7692	-0.8452	-0.8980	-1.0274	-0.8378
α_{31}	1.3863	1.5637	1.4348	1.3926	1.3386	1.2438	1.3895

Table 3.4: The results of the simulations for the mixed logistic regression model (Model 3)

Parameter	True value	Upper extreme	Upper quartile	Median	Lower quartile	Lower extreme	Average
Initial values set as the true values							
β_{11}	-2.1129	-0.2943	-1.5967	-2.2683	-2.8381	-4.5911	-2.2618
β_{12}	1.6057	3.3711	2.1703	1.6764	1.2100	-0.1373	1.7061
β_{21}	-0.9692	0.0446	-0.8472	-1.0817	-1.4547	-2.1604	-1.1588
β_{22}	1.3805	2.6180	1.7743	1.4772	1.1977	0.3613	1.5383
α_{11}	-1.2962	-0.4551	-0.9429	-1.1646	-1.5292	-2.1857	-1.2114
α_{12}	-0.4505	0.3340	-0.2551	-0.4819	-0.6523	-1.0381	-0.5067
α_{21}	-1.3148	-0.8501	-1.1829	-1.3497	-1.4191	-1.6918	-1.3164
α_{22}	1.0811	1.3219	1.1537	1.0830	0.9804	0.7596	1.0718
α_{31}	0.6973	1.0964	0.8138	0.6710	0.5802	0.3398	0.6881
α_{32}	0.7499	1.2727	0.9020	0.7549	0.5985	0.2325	0.7604
Initial values chosen step by step							
β_{11}	-2.1129	-0.2945	-1.5197	-2.1655	-2.8181	-4.6014	-2.2109
β_{12}	1.6057	3.3687	2.1429	1.6063	1.1472	-0.3411	1.6660
β_{21}	-0.9692	0.0457	-0.8123	-1.0815	-1.4438	-2.3682	-1.1286
β_{22}	1.3805	2.5867	1.7570	1.4766	1.1845	0.3604	1.5061
α_{11}	-1.2962	-0.4557	-0.9293	-1.1561	-1.5302	-2.1914	-1.2054
α_{12}	-0.4505	0.3340	-0.2544	-0.4896	-0.6639	-1.0377	-0.5096
α_{21}	-1.3148	-0.8394	-1.1830	-1.3459	-1.4153	-1.6962	-1.2974
α_{22}	1.0811	1.3233	1.1550	1.0768	0.9769	0.7526	1.0577
α_{31}	0.6973	1.0967	0.8149	0.6848	0.5824	0.3398	0.7022
α_{32}	0.7499	1.2727	0.8991	0.7518	0.5948	0.1788	0.7455

Table 3.5: Number of trout with liver tumors/number in tank

Dose (ppm)	Aflatoxin B1	Aflatoxicol
0.010	3/86, 5/86, 4/88, 2/86	9/87, 5/86,2/89,9/85
0.025	14/87,14/90, 9/83 12/88	30/86, 41/86, 27/86, 34/88
0.050	29/90, 31/89, 33/89, 26/87	54/89, 53/86, 64/90, 55/88
0.100	44/86,40/80, 44/89, 43/88	71/88,73/89, 65/88, 72/90
0.250	62/87,67/88,59/88,58/84	66/86, 75/82, 72/81,73/89

Table 3.6: Logistic regression and mixed logistic regression model estimates for fish data.

# (j)	Mixing probability			Binomial parameters				log-likelihood[1]	AIC	BIC
	β_{j0}	β_{j1}	β_{j2}	α_{j0}	α_{j1}	α_{j2}	α_{j3}			
Logistic regression model (1-component)										
1	NA	NA	NA	-1.758 (0.0847)	11.93 (0.6750)	0.8911 (0.1143)	2.402 (1.154)	-1930.38	-1934.38	-1937.76
1	NA	NA	NA	-1.839 (0.0764)	12.82 (0.5424)	1.063 (0.0803)		-1932.62	-1935.62	-1938.15
2-component mixture										
1	-43.89	1170.6	-0.0103	-0.9220	7.4548	1.4198	-2.1632	-1757.48	-1768.48	-1776.77
2				-4.0710	90.50	0.1337	47.67			
1	-42.81	1141.58		-0.9220	7.4547	1.4198	-2.1631	-1757.48	-1767.48	-1775.92
2				-4.0710	90.50	0.1337	47.67			
1	0.4095			-0.9232	7.4613	1.4141	-2.1311	-1784.36	-1793.36	-1800.96
2				-4.0708	90.4747	0.1356	47.48			
1	-42.72	1139.6		-0.8156	6.6201	1.1676		-1760.57	-1768.57	-1775.32
2				-4.7821	122.94	1.1716				
1	-44.38 (407.8)	1183.7 (4.453)		-0.8161 (0.0982)	6.6209 (0.6155)	1.1686 (0.0851)		-1760.57	-1767.57	-1773.48
2				-4.7798 (0.2896)	122.92 (12.49)					
1	-2.6492	64.73		-0.5843	7.9411			-1834.98	-1840.98	-1846.05
2				-3.8221	87.33					
1	4.9315			-1.5167				-1922.26	-1926.26	-1929.64
2	-100.94			0.7843						
3-component mixture										
1	25.66	-195.59	0.2586	-1.3636	13.58	1.1820	1.5813	-1750.80	-1768.80	-1784.00
2	74.16	-1500.7	-0.3769	-4.0710	90.50	0.1337	47.67			
3				-0.6032	5.9736	1.6905	-3.5038			

[1] Log-likelihood does not include the constant term.

Table 3.7: Parameter estiamtes for four models for fish data

Covariates	Logistic regression	Quasi-likelihood I	Quasi-likelihood II	Mixed logistic regression	
				component 1	component 2
α_0 (intercept)	-1.839 (0.0764)	-1.839 (0.2434)	-1.841 (0.2424)	-0.8161 (0.0982)	-4.7798 (0.2896)
α_1 (dose level)	12.82 (0.5424)	12.82 (1.7278)	12.81 (1.714)	6.6209 (0.6155)	122.92 (12.49)
α_2 (treatment)	1.063 (0.0803)	1.063 (0.2558)	1.058 (0.2553)	1.1686 (0.0851)	
Unexplained variance	$\sigma^2=1.0$	$\sigma^2=10.15$	$\phi=0.1056$	NA	

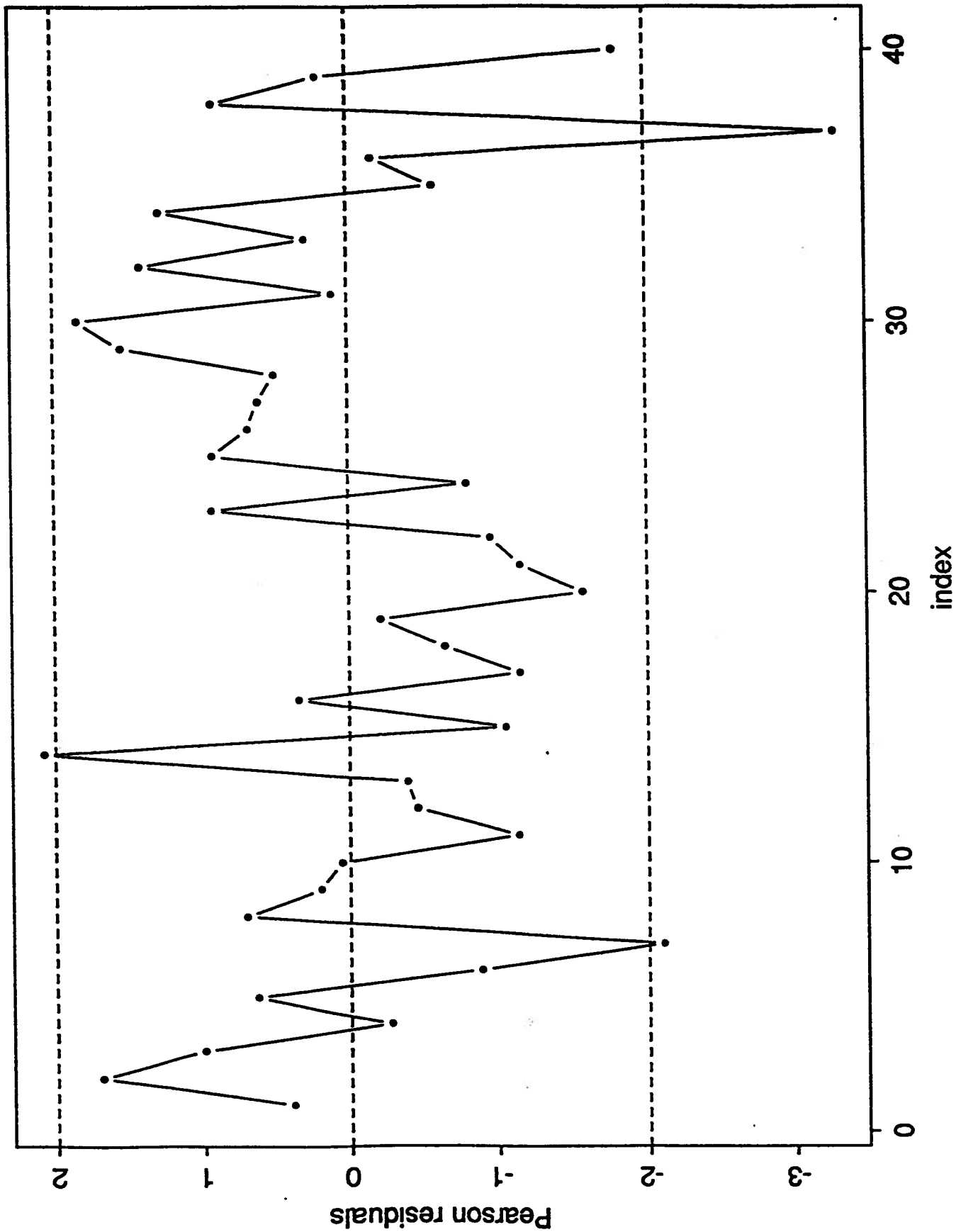

Figure 3.1: The index plot of the Pearson residuals from the fitted logistic regression model for the fish data.

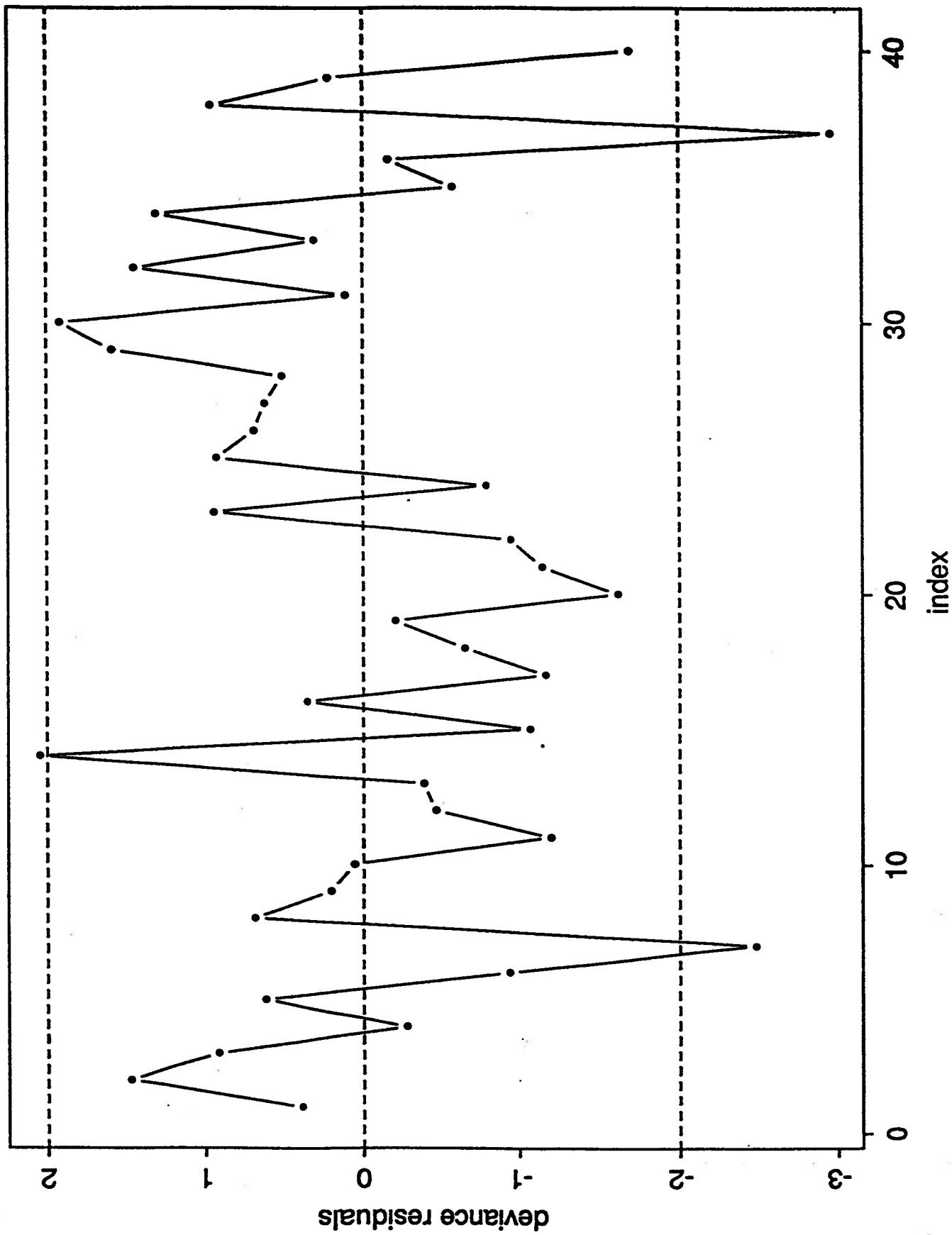

Figure 3.2: The index plot of the deviance residuals from the fitted mixed logistic regression model for the fish data.

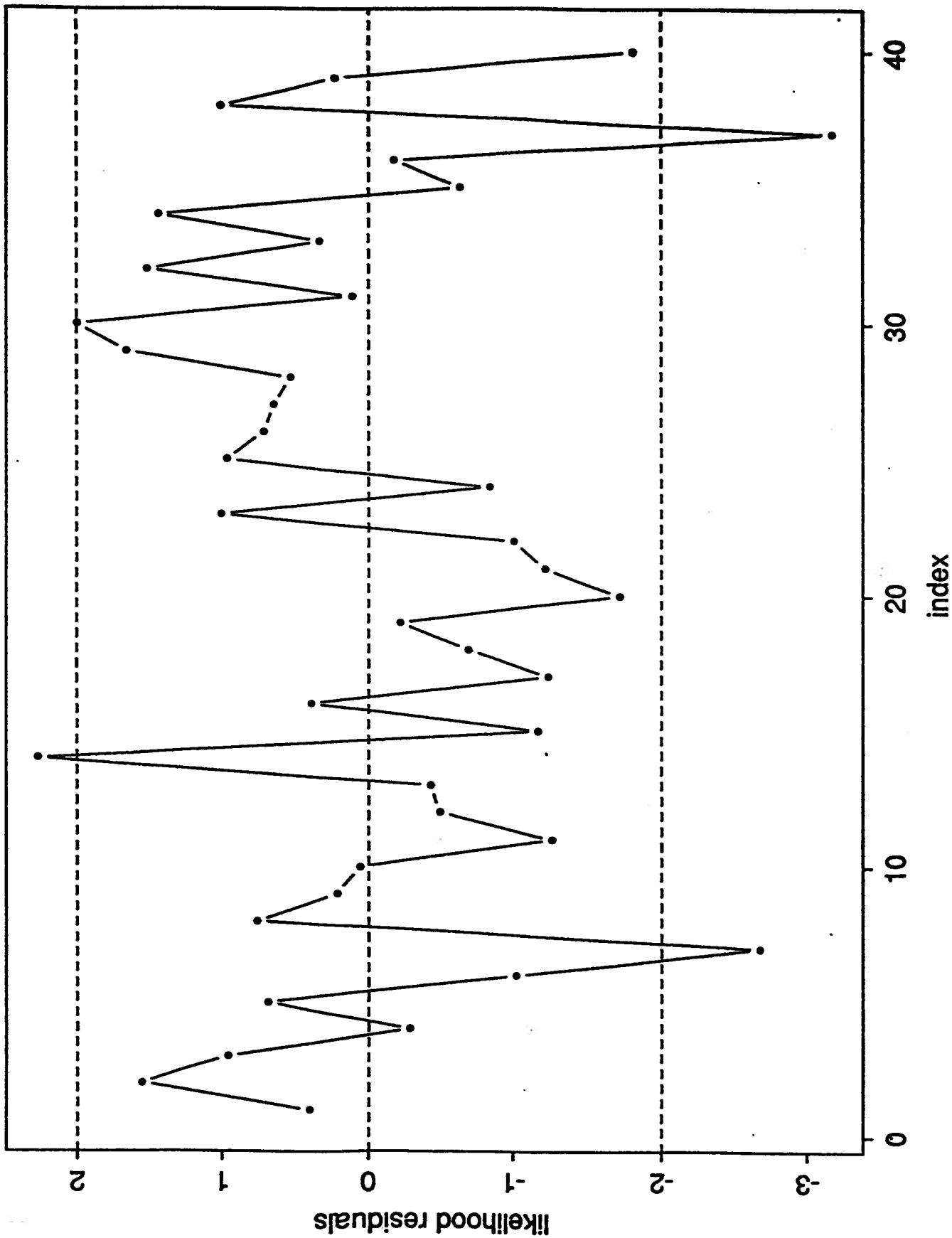

Figure 3.3: The index plot of the likelihood residuals from the fitted mixed logistic regression model for the fish data.

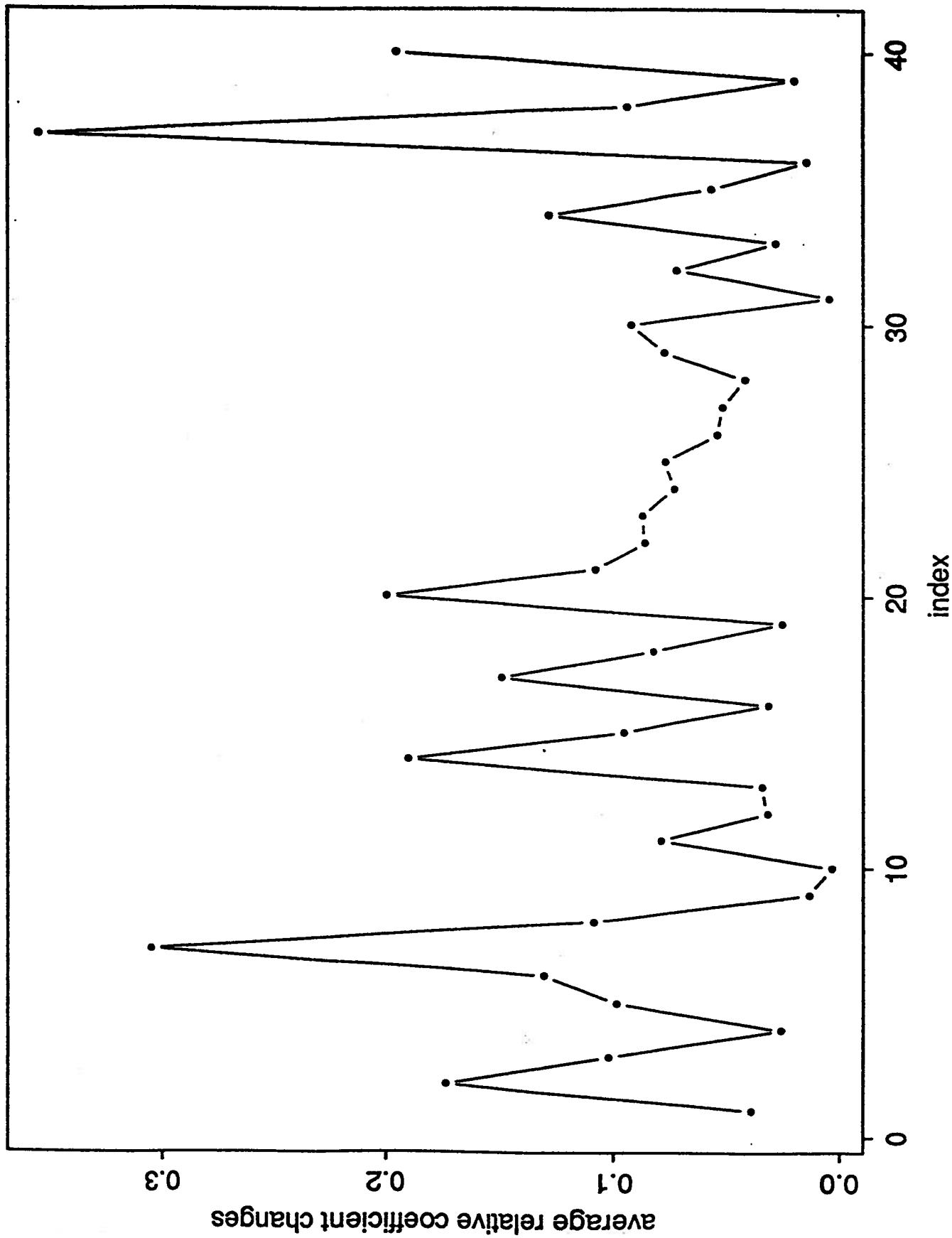

Figure 3.4: The index plot of the average relative coefficient changes from the fitted mixed logistic regression model for the fish data.

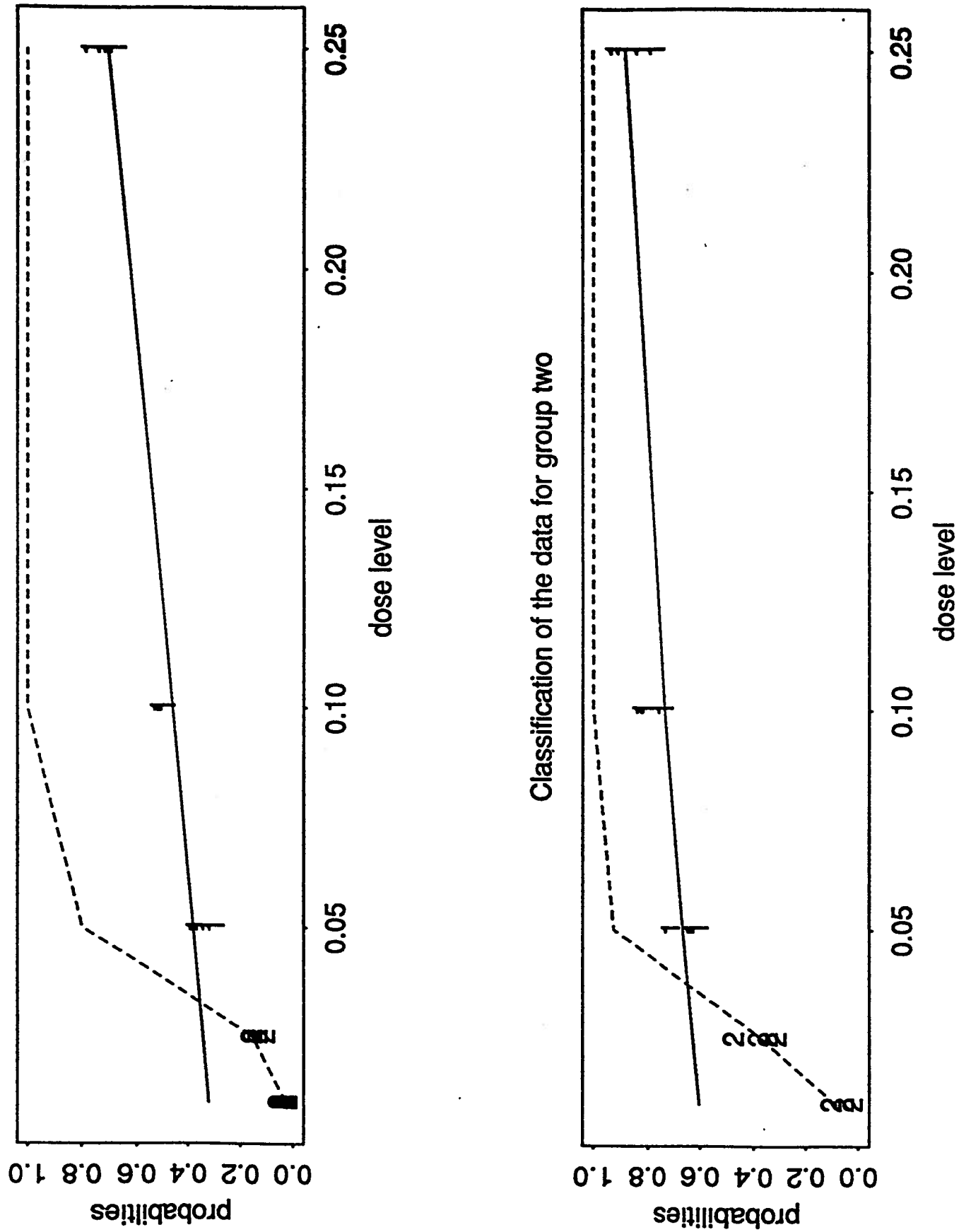

Figure 3.5: Classification and dose-response curves for fish data.

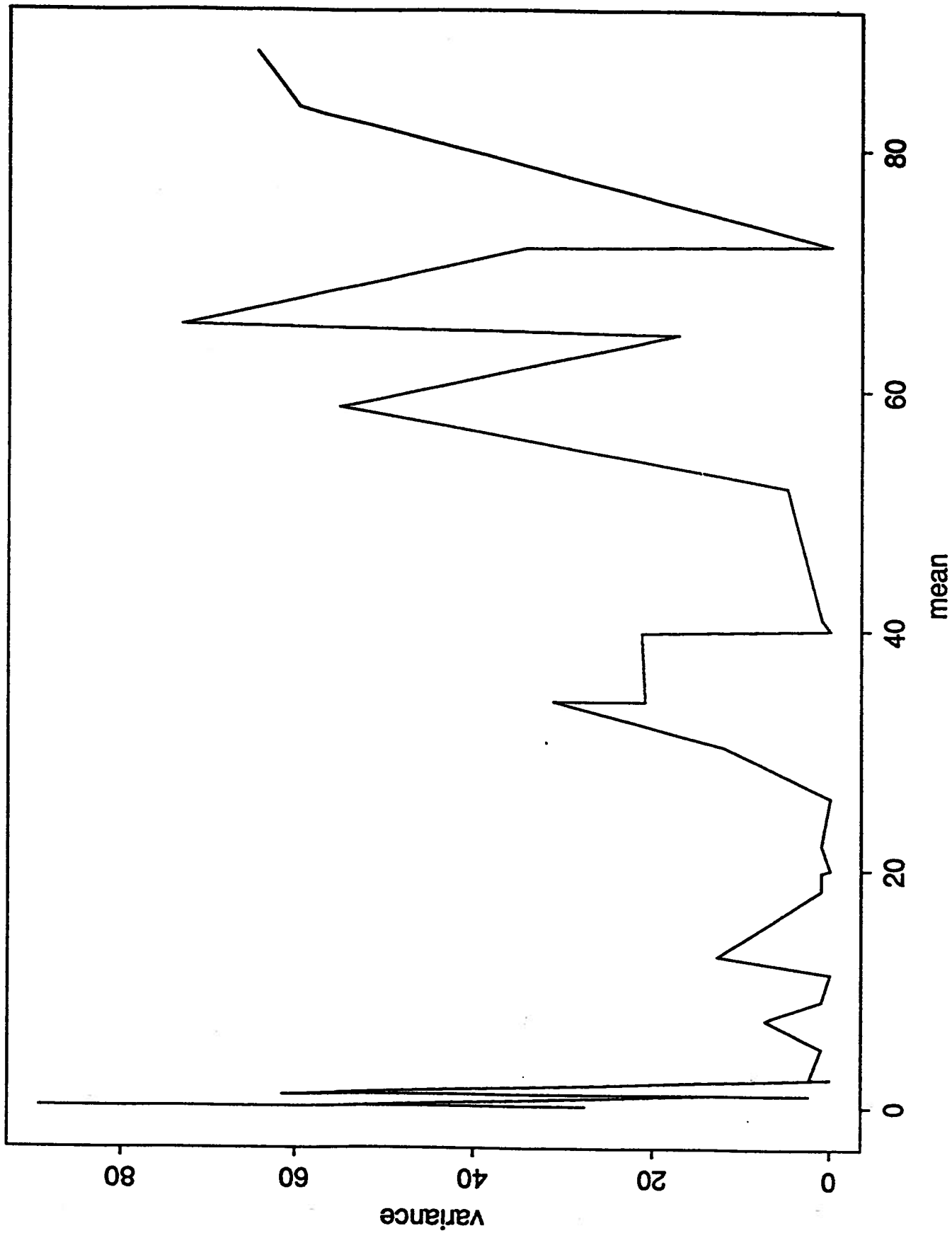

Figure 3.6: The plot of the mean-variance relationship based on the fitted mixed logistic regression model for the fish data.

Chapter 4

Summary, Conclusions and Future Research

In this chapter, we summarize similarities and differences between the mixed Poisson regression and mixed logistic regression models discussed in the previous chapters. Furthermore, we discuss some extensions of these mixed regression models and related remaining issues for future research. Section 4.2 formulates a mixed exponential family regression model which includes the mixed Poisson regression and mixed logistic regression models as special cases. Section 4.3 concerns a hidden Markov Poisson regression models for longitudinal data. We give some preliminary results of this model.

4.1 Summary and Conclusions

There are many similarities between the mixed Poisson regression and mixed logistic regression models discussed in Chapters 2 and 3. These are that

- both models assume an unobserved mixing process which can occupy any one of c states where c is finite and unknown; independent pairs of observed and unobserved random variables; covariates consisting of two parts: one related to the mixing probabilities, and the other to the component parameters; the same multinomial link in the mixing probabilities;

- both models can model overdispersion in the sense that the variances of the mixed regression models are larger than those specified by the mean-variance relationships of the corresponding usual regression models;

- parameters are estimated by maximum likelihood. Parameter estimates of both models are obtained by applying (1) the EM algorithm treating the unobserved random variable as missing data and (2) a quasi-Newton approach for the M-step and for maximizing the observed log likelihood functions;
- the model selection procedures for both models are the same, i.e., first determining the number of components by comparing the AIC and BIC values among the saturated models, and then carrying out inferences about regression parameters within c-component mixtures by likelihood ratio tests;
- classification, residual analysis and goodness-of-fit tests for both models are carried out in the same way.

There are several differences between the mixed Poisson regression and mixed logistic regression models. Obviously, the component distributions of the mixtures and link functions are different. This leads to different sufficient conditions for identifiability of these models. For the mixed Poisson regression models, the sufficient conditions for identifiability are virtually satisfied in all applications; for the mixed logistic regression models, since the sufficient conditions for identifiability depend on the binomial denominators, these may restrict the applications of these models in some cases. Although the algorithms for computation of parameter estimates for both models are similar, the implementation of these algorithms are quite different because there are different rescaling schemes to overcome numerical overflow or underflow problems. Note that coding these algorithms might be a formidable task.

Both the mixed Poisson regression and mixed logistic regression models provide new tools to analyze discrete data when data are overdispersed with respect to either the Poisson or binomial assumption. Allowing covariates in both mixing probabilities and the component parameters give a direct way to assess effects of each covariate on the

response variable. Using these models, we can classify observations into different groups characterized by different regression functions. This may give a more meaningful interpretation for overdispersion.

The mixed regression models are not always preferable to other models for modelling overdispersion such as parametric mixtures or quasi-likelihood regression. When overdispersion is reasonably modeled by a continuous mixing distribution, either parametric mixtures or quasi-likelihood regression models may be better. For the Poisson case, for instance, if extra-Poisson variation is caused by a random effect in the mean which is reasonably modelled by a continuous distribution, say a gamma distribution, then the negative binomial model is more suitable. Likewise, for the binomial case, if extra-binomial variation varies smoothly in the binomial denominators, Williams' quasi-likelihood models (1984) may be better. Nevertheless, the mixed regression models are suitable in many applications, which we have demonstrated in the previous chapters. The same technique of accommodating heterogeneity with mixture models can be applied to other cases. We discuss some generalizations below.

4.2 Mixed Exponential Regression Models

For a given one-parameter one-dimensional exponential model, the mean-variance relationship is determined by a single parameter. The one-parameter exponential density is

$$h(y)\exp(\theta y - \chi(\theta)),$$

where $h(y)$ is a real function, $\chi(\theta)$ is the log moment generating function with mean $\chi'(\theta) = \mu$, and variance $\chi''(\theta)$. Sometimes samples are found to be either too heterogeneous or homogeneous to be explained by a one-parameter exponential model of models in the sense that the implicit mean-variance relationship in such a model is violated by

the data. If the sample variance is large compared with that predicted by inserting the sample mean into the mean-variance relationship, overdispersion occurs. On the other hand, if sample variance is small compared with that predicted by the mean-variance relationship, underdispersion occurs. In this section, we suggest a mixed exponential regression model to adjust for overdispersion in terms of the mean-variance relationship of the one-parameter exponential model.

Let the random variable Y_i denote the ith response variable, and let $\{(y_i, x_i),\ i = 1, \ldots, n\}$ denote observations where y_i are observed value of Y_i, and $x_i = (x_i^{(m)}, x_i^{(r)})'$ are k-dimensional covariate vectors associated with y_i. Note that $x_i^{(m)}$ and $x_i^{(r)}$ are k_1-dimensional and k_2-dimensional vectors corresponding to the regression part of mixing probabilities and component parameters respectively. Usually the first element of $x_i^{(m)}$ and $x_i^{(r)}$ is 1 corresponding to an intercept. Our mixed exponential regression model assumes that

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

(2) for each observed response y_i, there is an unobserved random variable, Θ_i, representing the component which generates y_i. Further, the (Y_i, Θ_i) are pairwisely independent;

(3) conditional on covariate $x_i^{(m)}$, Θ_i follows a discrete distribution with c points of support, $1, \ldots, c$, and $Pr(\Theta_i = j \mid x_i^{(m)}, \beta) = p_j(x_i^{(m)}, \beta)$ where $\sum_{j=1}^{c} p_j(x_i^{(m)}, \beta) = 1$ and $p_j(x_i^{(m)}, \beta)$ is defined by

$$\begin{aligned} p_j(x_i^{(m)}, \beta) &\equiv p_{ij} \\ &= \frac{\exp(\beta_j' x_i^{(m)})}{1 + \sum_{k=1}^{c-1} \exp(\beta_k' x_i^{(m)})} \quad \text{for } j = 1, \ldots, c-1, \end{aligned} \tag{4.1}$$

and

$$\begin{aligned} p_{ic} &\equiv p_c(x_i^{(m)}, \beta) \\ &= 1 - \sum_{j=1}^{c-1} p_{ij} \end{aligned} \tag{4.2}$$

with $\beta = (\beta_1, \ldots, \beta_{c-1})'$ and $\beta_j = (\beta_{j1}, \ldots, \beta_{jk_1})'$, $j = 1, \ldots, c-1$, are unknown parameters. In fact, conditional on $x_i^{(m)}$, Θ_i follows a multinomial distribution $(1, p_{i1}, \ldots, p_{ic})$. Note that β appears in each p_{ij} for $1 \leq j \leq c$;

(4) conditional on $\Theta_i = j$, Y_i follows an one-parameter exponential distribution which we denote by

$$\begin{aligned} Y_i &\sim f_j\left(y_i \mid x_i^{(r)}, \alpha_j\right) \\ &= \exp(\theta_{ij} y_i - \chi(\theta_{ij})) \end{aligned} \tag{4.3}$$

where

$$\theta_{ij} \equiv h(x_i^{(r)}, \alpha_j) \quad \text{for } j = 1, \ldots, c,$$

where $\alpha \equiv (\alpha_1, \ldots, \alpha_c)'$ are unknown parameters , where $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jk_2})'$, $j = 1, \ldots, c$. Note that the component parameter θ_{ij} relate to covariates $x_i^{(r)}$ through the link function h.

Under the above assumptions the probability "density" of Y_i satisfies

$$f(y \mid x_i^{(r)}, x_i^{(m)}, \alpha, \beta) = \sum_{j=1}^{c} p_{ij} \exp(\theta_{ij} y_i - \chi(\theta_{ij})) \tag{4.4}$$

where p_{ij} and θ_{ij} are specified by (4.1),(4.2) and (4.3) respectively.

Note that the mixed Poisson regression and mixed logistic regression models discussed in the previous chapters are special cases of the mixed exponential regression models in which the component distributions are Poisson and binomial distributions respectively.

Another example of the mixed exponential regression models is the mixed normal regression model which assumes that the component distributions are normal distributions with common variance for all components. In this case, the component distributions can be denoted by

$$f_j\left(y_i \mid x_i^{(r)}, \alpha_j\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y_i - \mu_{ij})^2)$$

where

$$\mu_{ij} \equiv \alpha_j{}' x_i^{(r)} \quad \text{for } j = 1, \ldots, c,$$

Note that the link function is the identity function.

To apply the mixed exponential regression models, we need to show under what conditions the unconditional variance of Y is larger than that allowed by the one-parameter exponential distribution. The results given by Shaked (1980) may provide insight it. Since the different assumptions about the component distributions may lead to different conditions for identifiability of the mixed regression models, as we show in the previous chapters, we also need to show under what conditions the mixed exponential regression models is identifiable.

As we did for the mixed Poisson regression and mixed binomial regression models, parameter estimation of the mixed exponential regression model can be carried out by maximum likelihood. Furthermore, to obtain the maximum likelihood estimates requires using an iterative algorithm similar to ones in the previous chapters. Specifically, for a fixed number of components c, we may apply the EM algorithm by defining the unobserved random variable as missing data and using a quasi-Newton approach for the M-step. When either the observed likelihood or the parameter estimates do not change more than a given tolerance, we apply a quasi-Newton approach for maximizing the observed likelihood function.

After fitting data to the proposed model, we need to carry out residual analysis to identify possible outliers and influential observations and goodness-of-fit test for the fitted model. As we do for the mixed Poisson and logistic regression models, we propose using Pearson, deviance and likelihood residuals as well as relative average coefficient changes in a similar way for this purpose. We also suggest using the estimated posterior probabilities from the fitted model to classify observations into c groups, each characterized by a regression function.

4.3 Hidden Markov Poisson Regression Models

In this section, we consider a statistical method for longitudinal discrete data where the objective of data analysis is to describe an observed count, y_{ki}, for subject k during the ith time interval Δt_i as a function of covariates, x_{ki}. Longitudinal data are characterized by the fact that there may exist some dependence structure between repeated observations for a subject. The model which we have developed assumes that the dependence between repeated observations for a subject is determined by a finite state Markov chain in such a way that conditional on a state, an observed count, y_{ki}, follows a Poisson distribution with mean specified by the product of exposure, $\Delta t_i = t_i - t_{i-1}$, and Poisson rate defined by a log linear function of covariates, x_{ki}, in which coefficients may vary from state to state. This model allows for overdispersion relative to the usual Poisson regression model.

Our initial motivation comes from economic studies which investigate the relationship between research and development and patent activity at firm level based on longitudinal discrete data associated with covariates. The previous studies have suggested that the data may be overdispersed relative to the usual Poisson regression and that there may exist some correlation between repeated observations for a firm (Hausman, Hall and Griliches (1984) and Hall, Griliches and Hausman (1986)). However these studies

have no discussion about directly modeling the dependence structure between repeated observations. Our approach explicitly specifies the dependence structure as a finite state Markov chain and estimates both the parameters of the Markov chain and coefficients in the Poisson regression corresponding to each underlying state.

In the context of generalized linear models, several approaches have been developed for longitudinal data. Liang and Zeger (1986) proposed a general framework for analysis of longitudinal data based on generalized linear models, and Zeger (1988), Kaufmann (1987), Stiratelli, Laird and Ware (1984), Zeger, Liang and Self (1985) and Zeger and Qaqish (1988) developed methods for serially correlated discrete observations. In applications to economics in which data are primarily continuous, some approaches allow parameter values suddenly to change according to the states of a Markov chain, c.f. Goldfeld and Quandt (1973), Lindgren (1978), Sclove (1983) and Tyssedal and Tjostheim (1988).

In applications without covariates, Albert (1992) proposed a two- state Markov mixture model for longitudinal epileptic seizure counts. Leroux and Puterman (1992), Leroux (1989) and Le, Leroux and Puterman (1992) developed a finite state Markov mixture model for the sequence of counts of fetal movements. Our approach extends their approaches by incorporating covariates into the model and allowing variable exposure. We also use a rescale scheme to overcome either over or under numerical flow in applying the EM algorithm so that our algorithm improves the ones proposed by these authors.

4.3.1 The Model

The model we study in this paper embeds a finite state Markov chain in Poisson regression in which the regression coefficients depend on the chosen state. Specifically, let $\{(y_{ki}, x_{ki}, t_{ki}); i = 1, \ldots, n_k, 0 = t_{k0} < t_{k1} < \ldots < t_{kn_k}\}$ be a sequence of observed data for a subject k, where y_{ki} is an observed count associated with covariates x_{ki} of d-dimension

during a time interval $\Delta t_{ki} = t_{ki} - t_{ki-1}$. For simplicity we suppress the subscript for subjects in the following discussion. A Markov Poisson regression model assumes

(1) The unobserved stochastic process has c possible states where c is finite and unknown;

(2) For each observed count, y_i, at time point t_i, there exists an unobserved discrete random variable, S_i, representing a state at which y_i is generated. Further, S_i has c points of support, $\{1, \ldots, c\}$;

(3) The S-process, $\{S_1, S_2, \ldots, S_n\}$, follows a c-state Markov chain with transition probabilities defined by

$$Pr(S_i = j \mid S_{i-1} = k) = p_{jk}, \quad j, k = 1, \ldots, c; \tag{4.5}$$

(4) Conditional on $S_i = j$, Y_i follows a Poisson distribution which we denote as

$$\begin{aligned} f_j\left(y_i \mid x_i, \Delta t_i, \alpha_j, S_i\right) &\equiv \mathrm{Po}\left(y_i \mid \lambda_{ij}\right) \\ &= \frac{1}{y_i!}[\lambda_j(x_i, \alpha_j)\Delta t_i]^{y_i} \exp\left[-\lambda_j(x_i, \alpha_j)\Delta t_i\right] \end{aligned} \tag{4.6}$$

where $y_i = 0, 1, \ldots,$ $\lambda_{ij} \equiv \Delta t_i \lambda_j\left(x_i, \alpha_j\right)$ and $\lambda_j\left(x_i, \alpha_j\right)$is a nonnegative function equal to the Poisson rate; for example,

$$\lambda_j\left(x_i, \alpha_j\right) = \exp\left(\alpha_j' x_i\right),$$

where $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jk_1})$, $j = 1, \ldots, c$, are unknown parameter vectors. Note that $\Delta t_i = t_i - t_{i-1}$ may equal 1 for all i or correspond to time of observation in time series data.

The above assumptions define a semi-Markov process $\{(Y_i, S_i); i = 1, \ldots, n, 0 = t_0 < t_1 < \ldots < t_n\}$ in which the transitions of the S-process follow a stationary, first-order,

Markov process, and the count, Y_i, is renewed at each transition point, t_i, so that the conditional component distributions for the count depend only upon which state is exited. Note that the Poisson rates of the conditional component distributions vary between states by different coefficients in the same Poisson regression specification. Furthermore, since the covariates can include parts of an individual's past history, the proposed model provides a means of relaxing the assumptions about the transition process of the renewal counts. Note that the transition probabilities p_{ij} do not depend on covariates.

Under the above assumptions, the joint probability "density" function of a sequence of observed counts, $Y = \{y_1, \ldots, y_n\}$, associated with covariates, $X = \{x_1, \ldots, x_n\}$, and exposure, $T = \{\Delta t_1, \ldots, \Delta t_n\}$, satisfies the following equation.

$$\begin{aligned} f(Y \mid X, T, \theta) &= \sum_{j=1}^{c} \sum_{S_2}^{c} \cdots \sum_{S_n=1}^{c} p_j^{(1)} f_j(y_1 \mid x_1, \Delta t_1, \alpha_j, S_1) \\ &\quad \prod_{i=2}^{n} p_{S_{i-1}S_i} f_{S_i}(y_i \mid x_i, \Delta t_i, \alpha_{S_i}, S_i) \end{aligned} \tag{4.7}$$

where $\theta = (\alpha_{11}, \ldots, \alpha_{1d}, \alpha_{21}, \ldots, \alpha_{2d}, \ldots, \alpha_{c1}, \ldots, \alpha_{cd}, p_{11}, \ldots, p_{1c}, \ldots, p_{c1}, \ldots, p_{cc})$ is an unknown parameter vector, $p_j^{(1)} = Pr(S_1 = j)$, $j = 1, \ldots, c$, are the probabilities of the initial states for the subject, $p_{S_{i-1}S_i}$ and $f_{S_i}(y_i \mid x_i, \Delta t_i, \alpha_{S_i}, S_i)$ are defined by (4.5) and (4.6) respectively.

Note that the probabilities of the initial states, $p_j^{(1)}$, are assumed known. We will discuss how to determine their values below. Note also that $\sum_{k=1}^{c} p_{jk} = 1$ for all j.

Some previously studied models are special cases of the above model.

- Choosing $c = 1$ yields a Poisson regression model;
- Choosing the transition probability matrix as an identity matrix yields an independent mixed Poisson regression model which is a special case of the generalized

mixed Poisson regression models discussed by Wang, Puterman, Le and Cockburn (1994);

- Setting $x_i = (1)$ and $t_i = 1$ for all i yields Markov Poisson mixture without covariates which is studied by Leroux and Puterman (1992) and Albert (1992).

4.3.2 Moment Structure

From the above definition we can derive the basic moment structure of observed counts. Using the properties of conditional expectation, we obtain

$$E(Y_i \mid S_i) = \lambda_{iS_i} \quad \text{and} \quad Var(Y_i \mid S_i) = \lambda_{iS_i}.$$

Thus the unconditional mean and variance of Y_i are

$$\begin{aligned} E(Y_i) &= E(E(Y_i \mid S_i)) = \sum_{j=1}^{c} Pr(S_i = j)\lambda_{ij} \qquad (4.8) \\ Var(Y_i) &= E(Var(Y_i \mid S_i)) + Var(E(Y_i \mid S_i)) \\ &= \sum_{j=1}^{c} Pr(S_i = j)\lambda_{ij} + \left\{ \sum_{j=1}^{c} Pr(S_i = j)\lambda_{ij}^2 - \{\sum_{j=1}^{c} Pr(S_i = j)\lambda_{ij}\}^2 \right\} \qquad (4.9) \end{aligned}$$

Since the second term in (4.9) is always nonnegative, (4.8) and (4.9) show that the proposed model can accommodate overdispersion relative to Poisson regression, and that the observed data are homogeneous if and only if $\lambda_{i1} = \ldots = \lambda_{ic}$ for all i.

The covariance of Y_i and Y_{i+m} is given by

$$\begin{aligned} cov(Y_i, Y_{i+m}) &= cov(E(Y_i \mid S), E(Y_{i+m} \mid S)) \\ &= E(\lambda_{iS_i}\lambda_{i+ms_{i+m}}) - E(\lambda_{iS_i})E(\lambda_{i+ms_{i+m}}) \\ &= \sum_{j=1}^{c}\sum_{k=1}^{c} \lambda_{ij}\lambda_{i+mk} Pr(S_i = j, S_{i+m} = k) - E(Y_i)E(Y_{i+m}). \end{aligned}$$

4.3.3 Identifiability

Along with the applications of the Markov Poisson regression models we must be concerned with the identifiability for the models. Without covariates, Teicher (1961, 1967) proves that both the class of finite Poisson mixtures and the class of all mixtures of Poisson distribution products are identifiable. We will apply these results to derive the sufficient conditions for identifiability for the model. But we first define identifiability for the Markov Poisson regression model as follows.

Definition: Consider the class of probability models, $\{f(Y \mid X, T, \theta)\}$, with $f(Y \mid X, T, \theta)$ defined by (4.7), a restriction that $\lambda_{11} < \ldots < \lambda_{1c}$, parameter space $\mathcal{C} \times \Theta$, sample space $\mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n$ and fixed covariate matrices X and T. The class of probability models is *identifiable* if for (c, θ), $(c^*, \theta^*) \in \mathcal{C} \times \Theta$,

$$f(Y \mid X, T, \theta) = f(Y \mid X, T, \theta^*) \tag{4.10}$$

for all $(y_1, \ldots, y_n) \in \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n$, implies $(c, \theta) = (c^*, \theta^*)$.

Note that the order restriction in the definition indicates that two models are equivalent if they agree up to permutations of parameters. We now provide a sufficient condition for identifiability as follows.

Theorem 3: The hidden Markov Poisson regression model is identifiable if the design matrix X is full rank.

Proof: Suppose that (c, θ) and (c^*, θ^*) satisfy (4.10), then summing up both sides of equation (4.10) for $y_2, \ldots, y_n$ respectively yields

$$\sum_{j=1}^{c} p_j^{(1)} \mathrm{Po}(y_i \mid \lambda_{1j}) = \sum_{j=1}^{c^*} p_j^{*(1)} \mathrm{Po}(y_i \mid \lambda_{1j}^*) \tag{4.11}$$

for all $y_1 \in \mathcal{Y}_1$. Since each side of equation (4.11) may be regarded as a finite Poisson mixture without covariates, Teicher's result (1961) implies that

$$c = c^*, \quad p_j^{(1)} = p_j^{*(1)} > 0 \quad \text{and} \quad \lambda_{1j} = \lambda_{1j}^* \tag{4.12}$$

for $j = 1, \ldots, c$.

Now summing up both sides of (4.10) for $y_3, \ldots, y_n$ yields

$$\sum_{j=1}^{c}\sum_{k=1}^{c} p_j^{(1)} p_{jk} \mathrm{Po}(y_1 \mid \lambda_{1j}) \mathrm{Po}(y_2 \mid \lambda_{2k}) = \sum_{j=1}^{c}\sum_{k=1}^{c} p_j^{(1)} p_{jk}^{*} \mathrm{Po}(y_1 \mid \lambda_{1j}) \mathrm{Po}(y_2 \mid \lambda_{2k}^{*}) \quad (4.13)$$

for all $(y_1, y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$. Since each side of equation (4.13) may be regarded as a finite mixture of two Poisson distribution products without covariates, Teicher's result (1967) implies that

$$\begin{aligned} \lambda_{2j} &= \lambda_{2j}, \quad \text{for } j = 1, \ldots, c, \\ p_j^{(1)} p_{jk} &= p_j^{(1)} p_{jk}^{*}, \quad \text{for } j, k = 1, \ldots, c, \end{aligned} \quad (4.14)$$

or

$$p_{jk} = p_{jk}^{*}, \quad \text{for } j, k = 1, \ldots, c. \quad (4.15)$$

For each $i > 2$, summing up both sides of (4.10) for $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$ yields

$$\sum_{k=1}^{c}\left(\sum_{j=1}^{c}\cdots\sum_{s_{i-1}=1}^{c} p_{s_{i-1}k}\right)\mathrm{Po}(y_i \mid \lambda_{ik}) = \sum_{k=1}^{c}\left(\sum_{j=1}^{c}\cdots\sum_{s_{i-1}=1}^{c} p_{s_{i-1}k}\right)\mathrm{Po}(y_i \mid \lambda_{ik}^{*}) \quad (4.16)$$

for all $y_i \in \mathcal{Y}_i$. (4.16) implies that

$$\lambda_{ij} = \lambda_{ij}^{*}, \quad \text{for } i = 3, \ldots, n \text{ and } j = 1, \ldots, c. \quad (4.17)$$

From (4.12), (4.14) and (4.17) we obtain

$$\exp(\alpha_j' x_i) = \exp(\alpha_j^{*\prime} x_i) \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, c.$$

This is equivalent to

$$(\alpha_j - \alpha_j^{*})' x_i = 0, \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, c,$$

or

$$(\alpha_j - \alpha_j^{*})' X = 0, \quad \text{for } j = 1, \ldots, c. \quad (4.18)$$

Thus a sufficient condition for identifiability is that X is full rank, in which case (4.18) implies that $\alpha_j = \alpha_j^*$ for $j = 1, \ldots, c$. We can assume that this sufficient condition holds without loss of generality, since if it does not we can reparameterize the model accordingly. □

4.3.4 Estimation

The EM algorithm

In order to find the maximum likelihood estimates of the unknown parameters for the above model, we apply the EM algorithm (Dempster, Laird and Rubin, 1977), treating the unobservable state variable S_i as missing information. In doing so, we represent a complete data set by introducing the following indicator functions

$$zz(i,j,k) = \begin{cases} 1 & \text{if } S_{i-1} = k \text{ and } S_i = j \\ 0 & \text{otherwise;} \end{cases}$$

$$z(i,j) = \begin{cases} 1 & \text{if } S_i = j \\ 0 & \text{otherwise.} \end{cases}$$

Thus the log-likelihood of the complete data set, $\{(y_i, x_i, \Delta t_i, z(i,j), z(i,j,k)); i = 1, \ldots, n$ and $j, k = 1, \ldots, c\}$, with $\theta = \theta^0$ is

$$\begin{aligned} Q(\theta \mid \theta^0) &= \log p_{s_1}^{(1)} + \sum_{i=2}^{n}\sum_{j=1}^{c}\sum_{k=1}^{c} zz(i,j,k) \log p_{jk} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{c} z(i,j) \log f_{S_i}(y_i \mid x_i, \Delta t_i, \alpha_{S_i}, S_i) \\ &\equiv \log p_{s_1}^{(1)} + Q_1(\theta_1 \mid \theta^0) + Q_2(\theta_2 \mid \theta^0) \end{aligned}$$

where $\theta_1 = (p_{11}, \ldots, p_{1c}, \ldots, p_{c1}, \ldots, p_{cc})$, $\theta_2 = (\alpha_{11}, \ldots, \alpha_{1d}, \ldots, \alpha_{c1}, \ldots, \alpha_{cd})$,

$$Q_1(\theta_1 \mid \theta^0) = \sum_{i=2}^{n}\sum_{j=1}^{c}\sum_{k=1}^{c} zz(i,j,k) \log p_{jk} \quad \text{and}$$

$$Q_2(\theta_2 \mid \theta^0) = \sum_{i=1}^{n}\sum_{j=1}^{c} z(i,j)\log f_j(y_i \mid x_i, \Delta t_i, \alpha_j, S_i = j)$$

The EM algorithm finds the maximum likelihood estimates by proceeding iteratively in two steps: E-step and M step. At the E- step, it replaces the missing data in $Q(\theta \mid \theta^0)$ by its expectation, conditional on the observation data and the initial values of the parameters. At the M-step, it finds the estimates of the parameters by maximizing the expected log likelihood for the complete data set, conditional on the observed data. It repeats the two steps until the log likelihood of the observed data no longer increases. Note that the EM algorithm guarantees that the log likelihood does not decrease for each iteration.

In our case, the E-step of the EM algorithm updates the expected values of the missing data $z(i,j)$ and $zz(i,j,k)$ in each iteration, given the observed data and the initial values of the parameters. By definition,

$$\begin{aligned}
\hat{z}(i,j) &= E\{z(i,j) \mid y_1,\ldots,y_n\} \\
&= Pr(S_i = j \mid y_1,\ldots,y_n) \\
&= Pr(y_1,...,y_n, S_i = j)/\sum_k Pr(y_1,...,y_n, S_i = k) \\
\hat{zz}(i,j,k) &= E(z(i,j,k) \mid y_1,\ldots,y_n) \\
&= Pr(z(i,j,k) \mid y_i,\ldots,y_n) \\
&= Pr(y_1,\ldots,y_n, S_{i-1} = j, S_i = k)/Pr(y_1,...,y_n) \\
&= Pr(y_1,...y_{i-1}, S_{i-1} = j)p_{jk}Pr(y_i,...,y_n \mid S_i = k)
\end{aligned}$$

As first proposed by Baum et al. (1970), we use the following quantities to set up the forward-backward recursive formula for the computation of $\hat{z}(i,j)$ and $\hat{zz}(i,j,k)$,

$$\begin{aligned}
a_j(i) &= Pr(y_1,\ldots,y_i, S_i = j), \quad \text{for } i = 2,\ldots,n \text{ and} \\
a_j(1) &= p_j^{(1)} f_j(y_1 \mid x_1, \Delta t_1, \alpha_j, S_1 = j) \quad \text{for } j = 1,\ldots,c,
\end{aligned}$$

$$
\begin{aligned}
b_j(i) &= Pr(y_{i+1}, \ldots, y_n \mid S_i = j), \quad \text{for } i = 1, \ldots, n-1 \text{ and} \\
b_j(n) &= 1 \text{ for } j = 1, \ldots, c
\end{aligned}
$$

Thus $\hat{z}(i,j)$ and $\hat{zz}(i,j,k)$ can be written as

$$\hat{z}(i,j) = a_j(i)b_j(i) / \sum_{j=1}^{c} a_j(n) \tag{4.19}$$

$$\hat{zz}(i,j,k) = p_{jk} f_j(y_i \mid x_i, \Delta t_i, \alpha_j, S_i = j) a_j(i-1) b_k(i) / \sum_{j=1}^{c} a_j(n) \tag{4.20}$$

The advantage of the above expressions is that there are the following recursive formula to compute $a_j(i)$ and $b_j(i)$:

$$
\begin{aligned}
a_j(i) &= \sum_{k=1}^{c} Pr(y_1, ..., y_n, S_{i-1} = k, S_i = j) \\
&= \sum_{k=1}^{c} Pr(y_1, ..., y_n, S_{i-1} = k) p_{kj} f_j(y_i \mid x_i, \Delta t_i, \alpha_j, S_i = j) \\
&= \sum_{k=1}^{c} a_k(i-1) p_{kj} f_j(y_i \mid x_i, \Delta t_i, \alpha_j, S_i = j) \\
b_j(i) &= \sum_{k=1}^{c} Pr(y_{i+1}, ..., y_n, s_{i+1} = k \mid S_i = j) \\
&= \sum_{k=1}^{c} p_{jk} f_j(y_{i+1} \mid x_{i+1}, \Delta t_{i+1}, \alpha_j, S_i = j) Pr(y_{i+1}, ..., y_n \mid S_i = j) \\
&= \sum_{k=1}^{c} p_{jk} f_j(y_{i+1} \mid x_{i+1}, \Delta t_{i+1}, \alpha_j, S_i = j) b_k(i+1)
\end{aligned}
$$

The M-step is equivalent to maximizing the following two functions with respect to θ_1 and θ_2 separately:

$$
\begin{aligned}
\tilde{Q}_1(\theta_1 \mid \theta^0) &= \sum_{i=2}^{n} \sum_{j=1}^{c} \sum_{k=1}^{c} \hat{zz}(i,j,k) \log p_{jk} \quad \text{and} \\
\tilde{Q}_2(\theta_2 \mid \theta^0) &= \sum_{i=1}^{n} \sum_{j=1}^{c} \hat{z}(i,j) \log f_j(y_i \mid x_i, \Delta t_i, \alpha_j, S_i = j).
\end{aligned}
$$

To maximize $\tilde{Q}_1(\theta_1 \mid \theta^0)$ with respect to θ_1, the estimated values of the transition probabilities, $\hat{\theta}_1 = (\hat{p}_{jk})$, should satisfy the following equation

$$\frac{\partial \tilde{Q}_1}{\partial \theta_1} \mid_{\hat{\theta}_1} = 0. \tag{4.21}$$

Solving (4.21), we obtain $\hat{\theta}_1 = (\hat{p}_{jk})$ by

$$\hat{p}_{jk} = \frac{\sum_{i=2}^{n} \hat{z}z(i,j,k)}{\sum_{i=2}^{n} \sum_{l=1}^{c} \hat{z}z(i,j,k)}, \quad \text{j,k =1,...,c.} \tag{4.22}$$

To maximize $\tilde{Q}_2(\theta_2 \mid \theta^0)$ with respect to θ_2, the estimated value $\hat{\theta}_2$ should satisfy the following equation

$$\frac{\partial \tilde{Q}_2}{\partial \theta_2} \mid_{\hat{\theta}_2} = 0. \tag{4.23}$$

However there are usually no closed form for the solution of (4.23). We use the quasi-Newton approach (Nash, 1990) to solve it for $\hat{\theta}_2$.

We now summarize the EM algorithm for the hidden Markov Poisson regression model below.

Step 0: Specify starting values $\theta_1^{(0)}$ and $\theta_2^{(0)}$ and a tolerance ϵ;

Step 1: (E-step) Compute $\hat{z}(i,j)$ and $\hat{z}z_(i,j,k)$ using (4.19) and (4.20) respectively, for $i = 2, \ldots, n$ and $j, k = 1, \ldots, c$;

Step 2: (M-step)

1. Find the values of $\hat{\theta}_1 = \hat{p}_{jk}$ using (4.22);
2. Find the values of $\hat{\theta}_2$ to solve (4.23) using the quasi-Newton approach (Nash, 1990);

Step 3: If $\| \hat{\theta}_1 - \theta_1^{(0)} \| \equiv \sum_{j=1}^{c} \sum_{k=1}^{c} \mid \hat{p}_{jk} - p_{jk}^{(0)} \mid \geq \epsilon$ or $\| \hat{\theta}_2 - \theta_2^{(0)} \| \equiv \sum_{j=1}^{c} \sum_{k=1}^{d} \mid \hat{\alpha}_{jk} - \alpha_{jk}^{(0)} \mid \geq \epsilon$, set $\theta_1^{(0)} = \hat{\theta}_1$ and $\theta_2^{(0)} = \hat{\theta}_2$, and go to Step 1; Otherwise, stop.

The E-step of the EM algorithm

The difficulty to compute $\hat{z}(i,j)$ and $\hat{zz}(i,j,k)$ by (4.19) and (4.20) is that $a_j(i)$ and $b_j(i)$ converge to 0 or ∞ very fast as i increases. This will cause underflow problems in the computation. To overcome this difficulty, we introduce an approach to rescale $a_j(i)$ and $b_j(i)$ so that both maximum values are around 1 for each i. This approach takes the special structure of the model into account. It first represents, for such j that $a_j(i) \neq 0$,

$$\begin{aligned} a_j(i) &= (\sum_k^c a_k(i-1)p_{kj}) f_j(y_i \mid x_i, \Delta t_i, \alpha_j, S_1 = j) \\ &= \exp\left\{\log\left(\sum_k^c a_k(i-1)p_{kj}) + \log(f_j(y_i \mid x_i, \Delta t_i, \alpha_j, S_1 = j)\right)\right\} \\ &\equiv \exp(q_{ij}) \end{aligned}$$

where $q_{ij} = \log(\sum_k^c a_k(i-1)p_{kj}) + \log(f_j(y_i \mid x_i, \Delta t_i, \alpha_j, s_1 = j))$. It then rescales $a_j(i)$ by multiplying $\exp(-\dot{maxt}_i)$ for such j that $a_j(i) \neq 0$, where $maxt_i \equiv \max_k\{q_{ik}\}$, and stores the order of $a_j(i)$ by $powera(i) = \sum_{s=1}^i maxt_s$. This order will be used to calculate the orders of $\hat{z}(i,j)$ and $\hat{zz}(i,j,k)$. The same procedure is applied to calculate $b_j(i)$.

Before we state the computation of the E-step of the EM algorithm, we first define some notations for simplicity as follows:

$$\begin{aligned} f(i,j) &\equiv y_i! f_j(y_i \mid x_i, \Delta t_i, \alpha_j, S_i = j) \\ &= [\Delta t_i \lambda_j(x_i, \alpha_j)]^{y_i} \exp\left(-\Delta t_i \lambda_j(x_i, \alpha_j)\right) \\ &= \exp\{y_i \log(\Delta t_j) + y_i(\alpha_j{}' x_i) - \Delta t_i \exp(\alpha_j{}' x_i)\} \\ &\equiv \exp\left(r_{i,j}\right) \end{aligned}$$

where $r_{i,j} = y_i \log(\Delta t_j) + y_i(\alpha_j{}' x_i) - \Delta t_i \exp(\alpha_j{}' x_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, c$.

Note that factorials in the numerators and denominators of (4.19) and (4.20) are cancelled out. This simplifies the computation of the E-step.

The E-step of the EM algorithm can now be carried out as follows:

(a) compute $a_j(1) = p_j^{(1)} f(1,j)$, $j = 1, \ldots, c$, and set $powera(i) = 0$;

(b) compute $a_j(i)$ for $i = 2, \ldots, n$, and $j = 1, \ldots, c$ as follows:

1. identify an index set $K^a(i) \equiv \{k; \sum_{j=1}^{c} a_j(i-1)p_{jk} \neq 0, k = 1, \ldots, c\}$;
2. find $tempa(i) \equiv \max_{k \in K^a(i)} \{\log(\sum_{j=1}^{c} a_j(i-1)p_{jk}) + r_{i,k}\}$;
3. compute $a_j(i) = \exp\{\log(\sum_{k=1}^{c} a_k(i-1)p_{kj}) + r_{i,j} - tempa(i)\}$ for $j \in K^a(i)$ and 0 otherwise;
4. set $powera(i) = powera(i-1) + tempa(i)$.

(c) Set $b_j(n) = 1$, for $j = 1, \ldots, c$, and $powerb(n) = 0$;

(d) compute $b_j(n)$ for $i = n-1, n-2, \ldots, 1$ and $j = 1, \ldots, c$ as follows:

1. identify an index set $K^b(i) \equiv \{k; \sum_{j=1}^{c} p_{kj} b_j(i+1) \neq 0, k = 1, \ldots, c\}$;
2. find $tempb(i) \equiv \max_{k \in K^b(i)} \{\log(\sum_{j=1}^{c} p_{kj} b_j(i+1)) + r_{i+1,k}\}$;
3. compute $b_j(i) = \exp\{\log(\sum_{k=1}^{c} p_{jk} b_k(i-1)) + r_{i+1,j} - tempb(i)\}$ for $j \in K^b(i)$ and 0 otherwise;
4. set $powerb(i) = powerb(i-1) + tempb(i)$.

(e) For $i = 2, \ldots, n$ and $j = 1, \ldots, c$ compute $temp(i) \equiv \exp\{powera(i) + powerb(i) - powera(n)\}$ and

$$\hat{z}(i,j) = temp(i) a_j(i) b_j(i) / \sum_{j=1}^{c} a_j(n);$$

(f) For $i = 2, \ldots, n$ and $j = 1, \ldots, c$, compute

$$\hat{zz}(i,j,k) = temp(i) p_{jk} f(i,k) a_j(i-1) b_k(i) / \sum_{j=1}^{c} a_j(n).$$

4.3.5 The Probabilities of Initial States and Starting Values

In the above model we define the probabilities of initial states as known parameters. To determine their values, we consider two types of data: (1) data for a single subject and (2) data for several subjects. In the first case there is only the first observation directly related to the initial states so that there is little information about the initial probabilities. Thus we set $p_1^{(1)} = \ldots = p_c^{(1)}$. Since the data in this case usually contain a rather long sequence of observations, the values of the probabilities may not have significant effects on estimation in terms of asymptotic properties. Without covariates, Leroux (1989) proves that the effect of the probabilities vanishes as the number of observations increases.

In the second case we choose the values of the probabilities as the estimates of the mixing probabilities which are obtained by fitting the first observations of the subjects, $\{y_{1k}; k = 1, \ldots, m\}$, into a c-component mixed Poisson regression model with constant mixing probabilities and covariates in Poisson rates (Wang, Puterman, Le and Cockburn, 1993). Note that in this case the mixing probabilities can be equivalently interpreted as the the probabilities of initial states for the Markov mixture model. Further, in many applications like this, the data contain many subjects but short series.

To be able to run the EM algorithm, we need to choose the starting values for the unknown parameters in the model. The EM algorithm only guarantees, under some regularity conditions (Wu, 1983), that the parameter estimates are local maxima of the likelihood function. As the number of unknown parameters in the model increases, there may be more local maxima. Further, a poor choice of the starting values may slow down convergence with the EM algorithm. Indeed, in some cases where the likelihood is unbounded on the edge of parameter space, the sequence of estimates generated by the EM algorithm may diverge if the starting values are too close to the boundary. Hence for these reasons it is important to choose the starting values carefully so as to increase

the chance to achieve the maximum likelihood estimate. We use the following approach which works well in our applications.

We assume that c is known. We first fit the observed data into a c-component independent mixed Poisson regression model. Then we choose the initial values of the regression parameters as the corresponding estimates by the fitting. Further, we identify each observation with one of the c states if it has the largest value of the estimated posterior probabilities calculated by (4.19). We then calculate the frequencies of the transitions from state j to state k, and set these frequencies as the initial values of the corresponding transition probabilities, p_{ij}

4.3.6 Implementation and Remaining Issues

We suggest using BIC or AIC to determine the number of underlying states, and carrying out inference about parameters by likelihood ratio tests. Specifically, we first determine the number of components c by comparing BIC and AIC values among saturated models which include all covariates in Poisson means. After c is determined, we then carry out inference about regression parameters by likelihood ratio tests within c component mixture models. We will plan to conduct a Monte Carlo study to investigate this model selection procedure.

On the other hand, using the quantities $\hat{z}(i,j)$ and $\hat{zz}(i,j,k)$ from the fitted model, we can classify observations into one of c states, and identify transitions for each subject. This information may be useful in applications. Note that our code works well for fitting the fetal movement data (Leroux,1989) to the proposed model without covariates; the results are the same as those given by Leroux (1989).

Bibliography

[1] Aitkin, M., Anderson, D. and Hinde, J (1981), "Statistical Modelling of Data on Teaching Styles (with discussion)," *Journal of the Royal Statistical Society*, Ser. **A** 144, 419-461.

[2] Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," Second International Symposium on Information Theory, (B.N. Petrov and F. Csaki, Eds.), Budapest: Akademia Kaido, 267-81.

[3] Akaike, H. (1974), "A New Look At the Statistical Model Identification," *IEEE Trans. on Automatic Control*, **AC-19**, 716-23.

[4] Albert, P.S., (1991), "A two-state Markov model for a time series of epileptic seizure counts," *Biometrics*, **47**, 1371-1381.

[5] Amritage,P. (1957), "Studies in the variability of pock counts," *J. Hug., Camb.*, **55**, 564-581.

[6] Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, Second Edition, New York: Wiley.

[7] Anscombe,F.J. (1950), "Sampling theory of the negative binomial and logarithmic series distributions," *Biometrika*, **37**, 358-382.

[8] Aranda-Ordaz, F.J., (1981), "Quantal response analysis for a mixture of population," *Biometrics*, **28**, 981-988.

[9] Ashford, R. and Walker, P.J. (1972), "Quantal Response Analysis For a Mixture of Populations,",*Biometrics* **28**, 981-988.

[10] Backer, R.J. and Nelder, J.A., (1978), *The GLIM systems, release 3*, Oxford: Numerical Algorithms Group.

[11] Bartlett,M.S. (1936), "Some notes on insecticide tests in the laboratory and in the field," *J. R. Statist. Soc., Suppl.*, **3**, 185-194.

[12] Baum, L.E., Petrie, T., Soules, G., and Weiss, N., (1970), "A maximization technique occuring in the statistical analysis of probabilitic functions of Markov chains," *Annals of Mathematics Statistics*, **41**, 164-171.

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

[14] Bock, R.D. and Aitkin, M., (1981), "Marginal maximum likelihood estimation of item parameters: application of an EM algorithm,", *Psychometrika*, **46**, 443-459.

[15] Bound, J., Cummins, C., Griliches, Z., Hall, B.H., and Jaffe., A., (1984), "Who does R and D and who patents?" *National Bureau of Economic Research Working Paper No. 908, in Z. Griliches, ed., R and D, Patents, and Productivity*, (Chicago: University of Chicago Press), 21-54.

[16] Breslow, N. (1984), "Extra-Poisson Variation in Log-linear Models," *Applied Statistics,* **33**, 38-44.

[17] Breslow, N. (1990a), "Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-likelihood Methods," *Journal of the American Statistical Association*, **85**, 565-571.

[18] Breslow, N. (1990b), "Further Studies in Variability of Pock Counts," *Statistics in Medicine*, Vol.**9**, 615-626.

[19] Brillinger, D.R. (1986), "The Natural Variability of Vital Rates and Associated Statistics (with discussion)," *Biometrics*, **42**, 693-734.

[20] Busvine, J.R., (1938), "The toxicity of ethylene oxide to Calandra oryzae, C.C. Granaria, Tribolium Castaneum, and Cimex Lectualarius," *Biology*, **25**, 605-632.

[21] Cameron, A.C. and Trivedi, P.K. (1990), " Regression-Based Tests for Overdispersion in the Poisson Model", *Journal of Econometrics*, **46**, 347-364.

[22] Cameron, A.C. and Trivedi, P.K. (1986), " Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests," *Journal of Applied Econometrics*, **1**, 29-53.

[23] Carrol, R.J., Spiegelman, C.H., Lan, K.K.G., Bailey, K.T. and Abbott, R.D., (1984), "Errors-in-variables for binary regression models," *Biometrika*, **71**, 19-26.

[24] Collett, D., (1991), *Modelling Binary Data*, Champman & Hall.

[25] Collings,B.J. and Margolin,B.H. (1985), "Testing Goodness-of-Fit for the Poisson Assumption When Observations Are Not Identically Distributed," *Journal of the American Statistical Association*, **80**, 411-18.

[26] Cox, D.R. (1970), *The Analysis of Binary Data*, London: Chapman & Hall.

[27] Cox, D.R. (1983), "Some Remarks On Overdispersion," *Biometrika*, **70**, 269-274.

[28] Crowder, M.J. (1978), "Beta-Binomial Anova for Proportions," *Applied Statistics*, **27**, 34-37.

[29] Dean, C. and Lawless J.F. (1989), "Tests for Detecting Overdispersion in Poisson Regression Model," *Journal of the American Statistical Association*, **84**, 467-472.

[30] Dean, C.; Lawless J.F. and Willmot, G.E. (1989), "A mixed Poisson-inverse-Gaussian regression model," *Canadian Journal of Statistics*, **17**, 171-181.

[31] Dean, C. (1992), "Testing for Overdispersion in Poisson and Binormial Regression Models," *Journal of the American Statistical Association*, **87**, 451-457.

[32] 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* **B**, **39**, 1-38.

[33] Dennis, J.E. Jr. and Schanbel, R.B., (1983), *Numerical methods for unconstrained optimization and nonlinear equations*, Englewood Clifs, New Jersey: Prentice-Hall.

[34] Dietz, P.E.and Baker, S.P., (1974), "Drowning: epidemiology and prevention," *American Journal of Public Health*, **64**, 303-312.

[35] Duda, R.O. and Hart, P.E., (1973), *Pattern Classification and Scene Analysis*, New York: Wiley.

[36] Efron,E., (1986), "Double exponential families and their use in generalized linear regression". *Journal of the American Statistical Association*, **81**, 709-721

[37] Ehrenberg, (1972), *Repeated-Buying*, Amsterdam: North-Holland Publishing Co.; New York: American Elsevier Publishing Co.

[38] Everitt, B.S., and Hand, D.J., (1981), *Finite Mixture Distributions*, Chapman and Hall, Landon.

[39] Fienberg, S.E., (1981), *The Analysis of Cross-Classified Categorical Data*, Second Edition, Cambridge: The MIT press.

[40] Finney, D.J. (1976), "Radioligan assay," *Biometrics*,**32**, 721-740.

[41] Firth, D. (1987), "On the efficiency of quasi-likehood estimation," *Biometrika*,**74**, 233-245.

[42] Fisher,R.A. (1950), "The Significance of Deviations From Expectation in a Poisson Series," *Biometrics*,**6**, 17-24.

[43] Folks, J.L. and Chhikara, R.S., (1978), "The inverse Gaussian distribution and its statistical application–a review," *Journal of the Royal Statistical Society* **B**, **40**, 263-289.

[44] Follmann,D.A. and Lambert,D. (1989), "Generalizing Logistic Regression by Nonparametric Mixing," *Journal of the American Statistical Association*, **84**, 294-30.

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

[46] Formann, A.K. (1992), "Linear Logistic Latent Class Analysis for Polytomous Data," *Journal of American Statistical Association*, **87**, 476-486.

[47] Frome, E.L.; Kutner, M.H. and Beauchamp, J.J. (1973), "Regression analysis of Poisson-distributed data", *Journal of American Statistical Association*, **68**, 935-940.

[48] Frome, E.L. (1983), "The analysis of rates using Poisson regression models". *Biometrics*, **39**, 665-674.

[49] Fukunaga, K. (1972), *Introduction to Statistical Pattern Recognition*, New York: Academic Press.

[50] Ganio, L.M. and Schafer, D.W., (1992), "Diagnostics for overdispersion," *Journal of American Statistical Association*, **87**, 795-804.

[51] Ghosh, J.K. and Sen, P.K., (1985), "On the asymptotic performance of the log likelihod ratio statistic for the mixture model and related results," *Proc. Berkeley Conference in Jonor of Jerzy Neyman and Jack Kiefer (Vol. II)*, L.M. Le Cam and R.A. Olshen (Eds.). Monterey: Wadsworth, 789-806.

[52] Goldfeld, S.M. and Quandt, R.E., (1973), "A Markov model for switching regressions,", *Journal of Econometrics*, **1**, 3-16.

[53] Gourieroux, C., Monfort, A., and Trognon, A., (1984), "Pseudo maximum likelihood methods: applications to Poisson models," *Econometrica*, **52**, 701-720.

[54] Gram, L., (1988), "Experimental studies and controlled clinical testing of valproate and vigabatrin," *Acta. Neurol. Scand.,* **78**, 241-270.

[55] Griliches, Z., (1990), "Patent statistics as economic indicators: a survey," *Journal of Economic Literature,* **XXVIII**, 1661-1707.

[56] Guerrero, V.M. and Johnson, R.A., (1982), "Use of the Box-Cox transformation with binary response models," *Biometrika*, **69**, 309-314.

[57] Haberman, S.J., (1977), "Maximum likelihood estimation with incomplete data via the EM algorithm (discussion)," *Journal of the Royal Statistical Society* **B**, **39**, 1-38.

[58] Hall, B.H., Griliches, Z. and Hausman, J.A., (1986), "Patents and R and D: is there a lag," *International Economic Review*, **27**, 265-283.

[59] Hartigan, J.A. (1985a), "Statistical theory in clustering," *Journal of Classification*, **2**, 63-76.

[60] Hartigan, J.A. (1985b), "A failure of likelihood asymptotics for normal mixtures," *Proc. Berkeley Conference in Jonor of Jerzy Neyman and Jack Kiefer (Vol. II)*, L.M. Le Cam and R.A. Olshen (Eds.). Monterey: Wadsworth, 807-810.

[61] Hausman,J.A., Hall,B.H. and Griliches,Z., (1984), "Econometric models for count data with an application to the patents R and D relationship," *Econometrica*, **52**, 909-938.

[62] Hinde, J. (1982), "Compound Poisson regression model", *GLIM 82: Proc. Internat. Conf. Generalized Linear Models* (R. Gilchrist, ed.), Springer, Berlin, 109-121.

[63] Hill, J.R. and Tsai, C., (1988), "Calculating the efficiency of maximum quasi-likelihood estimation," *Applied Statistics*, **37**, 219-230.

[64] Hingson, R. and Howland, J., (1987), "Alcohol as a risk factor for jinjury or death resulting from accidnetal falls: a review of the literature," *J. Stud. Alc.*, **48**, 212-219.

[65] Holford, T.R. (1983), "The estimation of age, period and cohort effects for vital rates", *Biometrics*, **39**, 311-324.

[66] Hopkins, A., Davies, P. and Dobson, C., (1985), "Mathematical models of patterns of seizures," *Arch. Neurol.,* **42**, 463-467.

[67] Jorgensen, B. (1987), "Exponential dispersion models (with Discussion)," *Journal of the Royal Statistical Society* **B**, **49**, 127-162.

[68] Kaufmann, H. (1987), "Regression models for nonstationary categorical time series: asymptotic estimation theory," *Annal of Statistics*, **15**, 79-98.

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

[70] Lambert, D., and Roeder, K., (1993), "Overdispersion diagnostics for generalized linear models," working paper.

[71] Lawless, J.F. (1987a), "Regression Methods For Poisson Process Data," *Journal of the American Statistical Association*, **82**, 808-815.

[72] Lawless, J.F. (1987b), "Negative Binomial And Mixed Poisson Regression", *The Canadian Journal of Statistics*, **15**, 209-225.

[73] Le, N., Leroux, B.G. and Puterman, L.M. (1992), "Exact likelihood evaluation in a Markov mixture model for time series of seizure counts," *Biometrics*, **48**, 317-323.

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

[75] Leroux,B.G. (1989), "Maximum Likelihood Estimation for Mixture Distribution and Hidden Markov Models," University of British Columbia, Ph.D. dissertation.

[76] Leroux,B.G. and Puterman M.L. (1992), "Maximum Penalized Likelihood Estimation for Independent and Markov Dependent Mixture Models," *Biometrics*, **48**, 545-558.

[77] Liang, K.L. and Zeger, S.L., (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, **73**, 370- 384.

[78] Lindsay,B.G. (1983), "The Geometry of Mixing Likelihood: a General Theory," *The Annals of Statistics*, **11**, 86-94.

[79] Lindsay,B.G. and Roeder, K. (1992), "Residual diagnostics for mixture models," *Journal of the American Statistical Association*, **87**, 785-794.

[80] Linhart,H. and Zucchini,W. (1986), *Model Selection*, New York: John Wiley.

[81] Mannering, F.L. (1989), "Poisson analysis of commuter flexibility in changing routes and departure times," **Transpn. Res. B., 23B**, 53-60.

[82] Manton,K.G.; Woodbury,M.A., and Stallard, E., (1981), "A variance components approach to categorical data models with heterogeneous cell populations: Analysis of spatial gradients in lung cancer mortality rates in North Carolina counties," *Biometrics*, **37**, 259-269.

[83] Margolin, B.H.; Kaplan, N., and Zeiger,E., (1981), "Statistical analysis of the Ames salmonella/microsome test," *Proc. Nat. Acad. Sci. U.S.A.*, **76**, 3779-3783.

[84] 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.

[85] McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models* (Second Edition), London: Chapman and Hall.

[86] McDermott, F.T., (1977), "Alcohol, road crash casualties, and contermeasures," *A.N.Z. J. Surgery*, **47**, 156-161.

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

[88] Milton, J.G., Gotman, J., Remillard, G.M. and Adermann, F., (1987), "Timing of seizure recurrence in adult epileptic patients: a statistical analysis," *Epilepsia*, **28**, 471-478.

[89] Nash,J.C. (1990), *Compact Numerical Methods for Computers*, Adam Hilger.

[90] Neyman, J., (1959), "Optimal asymptotic tests of composite statistical hypotheses," *In Probability and Statistics,* Ed. U. Grenander, 213-234, New York: Wiley.

[91] Neuhaus, J.M., Kalbfleisch, J.D., and Hauck, W.W. (1991), "A comparison of cluster-specific and population averaged approaches for analyzing correlated binary data," *International Statistical Review*, **59**, 22-35.

[92] Ochi, Y. and Prentice, R.L., (1984), "Likelihood inference in a correlated probit regression model," *Biometrika*, **71**, 531-554.

[93] Otake, M. and Prentice, R.L., (1984), "The analysis of chromosomally aberrant cells based on beta-binomial distribution," *Radiation Research*, **98**, 456-470.

[94] Pierce, D.A. and Sands, B.R. (1975), "Extra-Bernoulli variation in binary data," *Technical Report*, **46**, Oregon State University, Department of Statistics.

[95] Pierce, D.A. and Schafer, D.W., (1986), "Residuals in generalized linear models," *Journal of the American Statistical Association* , **81**, 977-981.

[96] Pocock, S.J., Cook, D.G., and Beresford, S.A.A.,(1981), "Regression of area mortality rates on explanatory variables: What weighting is appropriate?," *Applied Statistics*, **30**, 370-384.

[97] Pregibon, D., (1981), "Logistic regression diagnostics," *Annals of Statistics,* **9**, 705-724.

[98] Prentice, R.L., (1976), "A generalization of the probit and logit methods for dose response curves," *Biometrics*, **32**, 761-768.

[99] Redner, R.A. and Walker, H.F., (1984), "Mixture densities, maximum likelihood and the EM algorithm", *SIAM*, **26**, 195-239.

[100] Roberts, H.V. (1991), *Data Analysis for Managers with Minitab*, The Scientific Press, South San Francisco.

[101] Schall, R. (1991), "Estimation In Generalized Linear Models With Random Effects", *Biometrika*, **78**, 719-27.

[102] Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, Vol.**6**, 461-464.

[103] Sclove, S., (1983), "Time-series segmentation: a model and a method," *Information Sciences*, **29**, 7-25.

[104] Shaked, M., (1980), "On mixtures from exponential families," *J.R. Statist. Soc. B*, **42**, 192-198.

[105] Simar,L. (1976), "Maximum Likelihood Estimation of a Compound Poisson Process," *The Annals of Statistics*, **4**, 1200-1209.

[106] Stein, G.Z., and Juritz,J.M. (1988), "Linear models with an inverse-Guassian distribution," *Comm. Statist. Theory Methods*, **17**, 557-571.

[107] Stiratelli, R., Laird, N. and Ware, J.H., (1984), "Random-effect models for series observations with binary response," *Biometrics*, **40**, 961-971.

[108] Tarone, R.E., (1976) "Testing the goodness of fit of the binomial distribution," *Biometrika*, **66**, 585-590.

[109] Teicher, H. (1961), "Identifiability of Mixtures," *Annals of Mathematical Statistics* **32**, 244-248.

[110] Teicher, H. (1963), "Identifiability of Finite Mixtures," *Annals of Mathematical Statistics* **34**, 1265-1269.

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

[112] Tweedie, M.C.K., (1957), "Statistical properties of inverse Gaussian distributions," *International Annals of Mathematical Statistics*, **28**, 362-372.

[113] Tyssedal, J.S., and Tjostheim, D., (1988), "An autoregression model with suddenly changing parameters and an application to stock market prices," *Applied Statistics* **37**, 353-369.

[114] Walker, P.J. (1966), "A Method of Measuring the Sensitivity of Trypanosome to Acriflavine and Trivalent Tryparsamide," *Journal of General Microbiol*, **43**, 45-58.

[115] Webb, G.R., Redman, S., Hennrikus, D.J., Kelman, G.R., Gibberd, R.W. and Sanson-Fisher, R.W., (1994), "The relationships between high-risk and problem drinking and the occurrence of work injuries and related absences," *Journal of Studies on Alcohol*, forthcoming.

[116] Wechsler, H., Kasey, E.H., Thum, D. and Demone, H.W., (1969), "Alcohol level and home accidents," *Public Health Reports*, **84**, 1043- 1050.

[117] Wedderburn, R.W.M. (1974), "Quasi-likelihood Functions, Generalized Linear Models and the Gauss-Newton Method," *Biometrika* **61**, 439-447.

[118] Wilesnsky, A.J., Ojemann, L. M., Temkin, N.R., Troupin, A.S. and Dodrill, C.B., (1981), "Clorazepate and phenobarbital as antiepileptic drugs: a double-blind study," *Neurology* **31**, 1271-1276.

[119] Williams, D.A. (1975), "The analysis of binary response from toxicological experiments involving reproduction and teratogenicity," *Biometrika*, **61**, 439-447.

[120] Williams, D.A. (1982), "Extra-binomial Variation in Logistic Linear Moldes," *Applied Statistics*, **31**, 144-148.

[121] Williams, D.A. (1984), "Generalized linear model diagnostics using deviance and single case deletions", *Applied Statistics*, **36**, 181-191.

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

[123] Zeger, S.L., (1988), "A regression model for time series of counts," *Biometrika*, **75**, 621-629.

[124] Zeger, S.L., Liang, K.Y. and Self, S.G., (1985), "The analysis of binary longitudinal data with time-independent covariates," *Biometrics*, **72**, 31-38.

[125] Zeger, S.L. and Qaqish, B., (1988), "Markov regression models for time series: a quasi-likelihood approach," *Biometrics*, **44**, 1019-1031.

Appendix A

1. Fortran program for computing the maximum likelihood estimates of the mixed Poisson regression model.

```
C*****************************************************************C
C
      PROGRAM GENMIX
C
C This code is designed for data in which each observation is associated
C with a time period. This code provides fitted values, Pearson's statistic,
C XSQR, Deviance, Pearson's and deviance residuals for the mixed model
C and for each component.
C NV = # OF VARIABLE REGRESSION COEFFICIENTS FOR EACH COMPONENT.
C NF = # OF COMMOM REGRESSION COEFFICIENTS.
C
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION BGUESS(30),OB(30),H(30,30),AGUESS(30),OA(30)
      DIMENSION PA(10),PB(10),PC(10),BT(30),DRES(1000),TDRES(1000)
      DIMENSION FIT(1000,12),RES(1000,6),std(30,30),  RESL(1000)
      dimension tobs(1000),ttime(1000),txm(1000,8),txm1(1000,8)
      dimension sign(1000),se(30),opar(30),w(1000)
      INTEGER N1,N2,IH1,IH2,MON,NEVALS,IFAIL,NSTEP
      INTEGER NT,NTEMP1,NTEMP2,IH3,NUM,ITER,NTOL
      OPEN( UNIT=1,FILE='lout1')
      OPEN( UNIT=2,FILE='residual')
      OPEN( UNIT=3,FILE='literes')
      OPEN( UNIT=7,FILE='result')
      open( unit=8,file='fitout')
      OPEN( UNIT=9,FILE='ldataout')
      READ(1,100) NOBS,NSTAT,NX,NX1,NV,NF
       write(*,100) nobs,nstat,nx,nx1,NV,NF
100   FORMAT(6I5)
      N1=(NSTAT-1)*NX
      N2=NSTAT*NV+NF
      NT=N1+N2
      NTEMP1=N1
      NTEMP2=N2
      READ(1,113) (OBS(I),TIME(I), I=1,NOBS)
110   FORMAT(F10.5)
       write(*,113) (obs(I),time(I),I=1,nobs)
      do 111 i=1,nobs
      READ (1,112) (XM(I,J),J=1,NX)
        if (i.gt.10) go to 111
       write(*,112) (xm(i,j),j=1,nx)
111   continue
112   FORMAT(6G16.8)
113   FORMAT(2F10.5)
      READ(1,110) (BGUESS(I),I=1,N1)
       write(9,110) (bguess(i),i=1,n1)
      DO 115 I=1,NOBS
      READ(1,112) (XM1(I,J),J=1,NX1)
       if (i.gt.10) go to 115
       write (*,112) (xm1(i,j),j=1,nx1)
115   CONTINUE
      READ(1,110) (AGUESS(I),I=1,N2)
       WRITE(9,110) (AGUESS(I),I=1,N2)
      do 118 i=1,nobs
      tobs(i)=obs(i)
      ttime(i)=time(i)
      do 116 j=1,nx
116   txm(i,j)=xm(i,j)
      do 117 j=1,nx1
117   txm1(i,j)=xm1(i,j)
118   continue
      NEVALS=1000
      IH1=N1
      IH2=N2
      IH3=NT
      MON=1
      TOL=0.0001
      TOLL=0.01
      DO 120 I=1,NOBS
120   TDRES(I)=0.0
      OBSINF=0.0
```

```
      DEV=0.0D0
      DO 150 I=1,NOBS
      IF (OBS(I).EQ.0.0) GO TO 150
      TDRES(I)=OBS(I)*(DLOG(OBS(I))-1.0D0D0)-obs(i)*dlog(time(i))
      DEV=DEV+TDRES(I)
      TEMP1=0.0D0
      NSTEP=INT(OBS(I))
      DO 145 J=1,NSTEP
      TEMP1=TEMP1+DLOG(DFLOAT(J))
  145 CONTINUE
      obsinf=obsinf+TEMP1-OBS(I)*DLOG(TIME(I))
  150 CONTINUE
      NTOL=NOBS
      ODEV=DEV
      OOBSINF=OBSINF
      do 155 i=1,ntol
  155 w(i)=0.0D0
  160 DO 888 ITER=0,NTOL
      PREL=-(10.0D0**10.0D0)
  200 DO 202 I=1,N1
  202 OB(I)=BGUESS(I)
      DO 205 I=1,N2
  205 OA(I)=AGUESS(I)
      CALL ESTEP(NTEMP1,NTEMP2,BGUESS,AGUESS)
      CALL MSTEP1(NTEMP2,AGUESS,H,P,IH2,NEVALS,IFAIL,MON)
      CALL MSTEP2(NTEMP1,BGUESS,H,P,IH1,NEVALS,IFAIL,MON)
      SSR1=0.0D0
      SSR2=0.0D0
      DO 350 I=1,N1
      SSR1=SSR1+(OB(I)-BGUESS(I))**2.0D0
  350  CONTINUE
      DO 352 I=1,N2
  352  SSR2=(OA(I)-AGUESS(I))**2.0D0+SSR2
      DO 353 K=1,N1
  353  BT(K)=BGUESS(K)
      DO 354 K=1,N2
  354  BT(K+N1)=AGUESS(K)
      CALL LLIKELY(NT,BT,F)
      TEMP=-F-PREL
      IF (TEMP.LT.TOLL) GO TO 359
      PREL=F
      IF (ITER.GT.0) GO TO 356
      f=-f-obsinF
      WRITE(9,1111) F
       write(*,1111) f
 1111 format(4x,G16.8)
  356  IF ((SSR1.GT.TOL).OR.(SSR2.GT.TOL))  GO TO 200
  359  call newton(nt,bt,h,p0,nt,nevals,ifail,mon,std)
      if (iter.eq.0) call Flikely(nt,bt,f,dreS)
      if (iter.gt.0) call llikely(nt,bt,f)
      do 364 i=1,n1
  364  bguess(i)=bt(i)
      do 365 i=1,n2
  365  aguess(i)=bt(n1+i)
      DEV=2*(DEV-(-F))
C      write(*,*) iter,odev,tdres(iter),f
      IF (ITER.EQ.0) TDEV=DEV
      IF (ITER.GT.0) GO TO 500
      f=-f-obsinf
      write(9,1111) f
      call gfit(n1,n2,bguess,aguess,XSQR,fit,RES,pa,pb,pc)
      do 366 i=1,nobS
      temp=fit(i,1)-fit(i,2)
      if (temp.eq.0.0D0)  sign(i)=0.0D0
      if (temp.ne.0.0D0)  sign(i)=temp/(abs(temp))
      dres(i)=(2*(tdres(i)-dres(i)))**(0.5D0D0)
      dres(i)=sign(i)*dres(i)
  366  continue
      write(9,4444)
 4444 format(4x,'goodness  of fit--XSQR, DeviancE')
      write(9,7777) XSQR, DEv
      do 368 i=1,nobs
      write(8,369)  (FIT(I,J),J=1,2+2*NSTAT)
c      WRITE(2,369)  (RES(I,J),J=1,1+NSTAT)
      write(7,112)  (z(i,j),j=1,nstat)
      NUM=1
      DO 367 J=1,NSTAT-1
      IF (Z(I,J).GT.Z(I,J+1))  GO TO 367
      NUM=J+1
  367  CONTINUE
      WRITE(2,370) FIT(I,1),FIT(I,2),RES(I,1),DRES(I),FIT(I,NUM+2),
```

```
c          RES(I,1+NUM),NUM
368  continue
369  format(12g16.8)
370  FORMAT(G12.6,x,g12.6,x,g12.6,x,g12.6,x,g12.6,x,g12.6,X,I2)
     DO 400 J=1,N1
400  BGUESS(J)=BT(J)
     DO 402 J=1,N2
402  AGUESS(J)=BT(J+N1)
     call estep(ntemp1,ntemp2,bguess,aguess)
     temp=dfloat(n1+n2)
     do 410 k=1,n1
     opar(k)=bguess(k)
     se(k)=(std(k,k)**(0.5D0D0))*temp
410  continue
     do 420 k=1,n2
     opar(k+n1)=aguess(k)
     se(k+n1)=(std(k+n1,k+n1)**(0.5D0D0))*temp
420  continue
     WRITE(9,5555)
     WRITE(9,7777) (BGUESS(I), std(I,i)**(0.5D0D0), I=1,N1)
     write(*,7777) (bguess(i),i=1,n1)
     write(9,6666)
     write(9,7777) (AGUESS(I), std(I+N1,i+n1)**(0.5D0D0),I=1,N2)
     write(*,7777) (aguess(i),i=1,n2)
     write(9,7787)
     write(9,7777) (pa(i),i=1,nstat)
     write(9,7797)
     write(9,7777) (pb(i),i=1,nstat)
     write(9,7799)
     write(9,7777) (pc(i),i=1,nstat)
     GO TO 504
500  RESL(ITER)=TDEV-DEV
     do 501 k=1,n1
     w(iter)=w(iter)+abs(bguess(k)-opar(k))/se(k)
     bguess(k)=opar(k)
501  continue
     do 502 k=1,n2
     w(iter)=w(iter)+abs(aguess(k)-opar(k+n1))/se(k+n1)
     aguess(k)=opar(k+n1)
502  continue
     IF (ITER.EQ.NTOL) GO TO 889
504  NOBS=NTOL-1
     DEV=ODEV-TDRES(ITER+1)
     if (iter.eq.0) go to 514
     do  510 k=1,iter
     obs(k)=tobs(k)
     time(k)=ttime(k)
     do 505 j=1,nx
505  xm(k,j)=txm(k,j)
     do 506 j=1,nx1
506  xm1(k,j)=txm1(k,j)
510  continue
     if (iter.eq.nobs) go to 888
514  DO 520 K=ITER+1,NOBS
     OBS(K)=TOBS(K+1)
     TIME(K)=tTIME(K+1)
     DO 515 J=1,NX
515  XM(K,J)=tXM(K+1,J)
     DO 516 J=1,NX1
516  XM1(K,J)=tXM1(K+1,J)
520  CONTINUE
888  CONTINUE
889  do 900 i=1,ntol
     resl(i)=sign(i)*(resl(i)**0.5D0D0)
     write(3,7778) dres(i),res(i,1),resl(i),w(i)
900  continue
5555 format(4x,'beta-vector')
6666 format(4x,'alpha-vector')
7777 format(4x,2g16.8)
7778 format(4x,4g16.8)
7787 format('pa')
7797 format('pb')
7799 format('pc')
9999 STOP
     END

     SUBROUTINE FUNCT(N,B,P)
C*************************************************************
C This subroutine computes the value of function Q1 in Chapter 2.
C Data input: N = dimension of vector B;
C             B = beta vector;
```

```
C     output: P = the function value Q1(B).
C*****************************************************************
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      INTEGER N
      DIMENSION B(N),BX(5)
      P=0.0D0
      DO 100 I=1,NOBS
      DO 8 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 6 M=1,NX
    6 BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
    8 CONTINUE
      P=P+Z(I,1)*BX(1)
      TEMP1=BX(1)
      IF (TEMP1.LT.0.0D0) TEMP1=0.0D0
      DO 20 J=2,NSTAT-1
      P=P+Z(I,J)*BX(J)
      IF (BX(J).GT.BX(J-1)) TEMP1=BX(J)
   20 CONTINUE
      P=P-TEMP1
      CALL AEXP(-TEMP1,TEMP2)
      DO 30 J=1,NSTAT-1
      CALL AEXP((BX(J)-TEMP1),TEMP3)
      TEMP2=TEMP2+TEMP3
   30 CONTINUE
      P=P-DLOG(TEMP2)
  100 CONTINUE
      P=-P
      RETURN
      END

      SUBROUTINE GRAD(N,B,G)
C*****************************************************************
C This subroutine computes the first derivative of Q1 (see eqn 2.21).
C Data input: N = dimension of vector B;
C             B = beta vector;
C     output: G = the derivative of Q1 at B.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      INTEGER N
      DIMENSION G(25),B(N),TEMP(25),BX(5)
      DO 5 I=1,N
    5 G(I)=0.0D0
      DO 100 I=1,NOBS
      DO 20 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
   10 BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
   20 CONTINUE
      TEMP1=BX(1)
      IF (TEMP1.LT.0.0D0) TEMP1=0.0D0
      DO 30 J=2,NSTAT-1
      IF (BX(J).GT.BX(J-1)) TEMP1=BX(J)
   30 CONTINUE
      CALL AEXP(-TEMP1,TEMP2)
      DO 40 J=1,NSTAT-1
      CALL AEXP((BX(J)-TEMP1),TEMP(J))
      TEMP2=TEMP2+TEMP(J)
   40 CONTINUE
      DO 60 J=1,NSTAT-1
      DO 50 M=1,NX
      G(M+(J-1)*NX)=G(M+(J-1)*NX)+XM(I,M)*(Z(I,J)-TEMP(J)/TEMP2)
   50 CONTINUE
   60 CONTINUE
  100 CONTINUE
      DO 200 I=1,N
  200 G(I)=-G(I)
      RETURN
      END

      SUBROUTINE MSTEP2(N,B,H,P0,IH,NEVALS,IFAIL,MON)
C*****************************************************************
C  This subroutine is a quasi-Newton algorithm (Nash, 1990) which
C  maximizes the function Q1.
```

```
C  Data input: N = dimension of vector B; B = beta vector;
C            IH = dimension of the Hessian matrix;
C            NEVALS = # of evaluations for the function Q1;
C     output: H = the Hessian matrix; P0 = maximum value;
C            B = optimal values of beta vector.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION B(N), H(IH,N)
      DIMENSION X(30), C(30), G(30), T(30)
      DOUBLE PRECISION K
      INTEGER COUNT
      DATA W,TOL/0.2,1.0D0D-4/,EPS/1.0D0D-6/
      IF (N.LT.0.OR.N.GT.23) GO TO 160
      IFN = N+1
      IG = 1
      RLIM=7.2D0*(10.0D0**74.0D0)
      CALL FUNCT(N,B,P0)
      IF(P0.GT.RLIM)GOTO180
      CALL GRAD(N,B,G)
C
C     RESET HESSIAN
C
   10 DO 30 I = 1,N
        DO 20 J = 1,N
   20     H(I,J) = 0.0D0
   30   H(I,I) = 1.0D0D0
      ILAST = IG
C
C     TOP OF ITERATION
C
   40 DO 50 I = 1,N
        X(I) = B(I)
   50   C(I) = G(I)
C
C     FIND SEARCH DIRECTION T
C
      D1 = 0.0D0
      SN=0.0D0
      DO 70 I = 1,N
        S = 0.0
        DO 60 J = 1,N
   60     S = S-H(I,J)*G(J)
        T(I) = S
        SN = SN+S*S
   70   D1 = D1-S*G(I)
C
C     CHECK IF DOWNHILL
C
      IF (D1.LE.0.0D0) GO TO 10
C
C     SEARCH ALONG T
C
      SN = 0.5D0D0/DSQRT(SN)
      K = DMIN1(1.0D0D0,SN)
   80 COUNT = 0
      DO 90 I = 1,N
        B(I) = X(I)+K*T(I)
        IF (DABS(B(I)-X(I)).LT.EPS) COUNT = COUNT+1
   90   CONTINUE
C
C     CHECK IF CONVERGED
C
      IF (COUNT.EQ.N) GO TO 150
      CALL FUNCT(N,B,P)
      IFN = IFN+1
      IF (IFN.GE.NEVALS) GO TO 170
      IF (P.LT.P0-D1*K*TOL) GO TO 100
      K = W*K
      GO TO 80
C
C     NEW LOWEST VALUE
C
  100 P0 = P
      IG = IG+1
      CALL GRAD(N,B,G)
      IFN = IFN+N
C
C     UPDATE HESSIAN
```

```
C
      D1 = 0.0D0
      DO 110 I = 1,N
        T(I) = K*T(I)
        C(I) = G(I)-C(I)
 110    D1 = D1+T(I)*C(I)
C
C     CHECK IF +VE DEF ADDITION
C
      IF (D1.LE.0.0D0) GO TO 10
      D2 = 0.0D0
      DO 130 I = 1,N
        S = 0.0D0
        DO 120 J = 1,N
 120      S = S+H(I,J)*C(J)
        X(I) = S
 130    D2 = D2+S*C(I)
      D2 = 1+D2/D1
      DO 140 I = 1,N
        DO 140 J = 1,N
 140    H(I,J) = H(I,J)-(T(I)*X(J)+T(J)*X(I)-D2*T(I)*T(J))/D1
      GO TO 40
 150 IFAIL = 0
C SUCCESSFUL CONCLUSION
      RETURN
 160 IFAIL = 1
C N OUT OF RANGE
      RETURN
 170 IFAIL = 2
C TOO MANY FUNCTION EVALUATIONS
      RETURN
 180 IFAIL = 3
C INITIAL POINT INFEASIBLE
      RETURN
2005 FORMAT( 2X,3G16.4)
      END


      SUBROUTINE AEXP(X,F)
C*****************************************************************
C This subroutine computes a exponential function value.
C Data input: X = real number;
C     output: F = exp(X).
C*****************************************************************
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      INTEGER NSTEP
      TEMP1=ABS(X)
      IF (TEMP1.GT.79.9D0) GO TO 50
      F=DEXP(X)
      GO TO 200
 50   IF (X.LT.-79.9D0) GO TO 150
      IF (X.GT.150.0D0) X=150.0D0
      F=1.0D0D0+X
      NSTEP=1
      FACTOR=1.0D0D0
      TEMP1=DFLOAT(NSTEP)
      TEMP2=X/TEMP1
 100  IF (TEMP2.LT.1.0D0D0) GO TO 200
      NSTEP=NSTEP+1
      TEMP1=DFLOAT(NSTEP)
      FACTOR=X/TEMP1
      TEMP2=TEMP2*FACTOR
      F=F+TEMP2
      GO TO 100
 150  F=0.0D0
 200  RETURN
      END


      SUBROUTINE FUNCT1(N,B,P)
C*****************************************************************
C This subroutine computes the value of function Q2 in Chapter 2.
C Data input: N = dimension of vector B;
C             B = alpha vector;
C     output: P = the function value Q2(B).
C*****************************************************************
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      INTEGER N
```

```
      DIMENSION B(N),BX(5)
      P=0.0D0
      DO 100 I=1,NOBS
      DO 8 J=1,NSTAT
      BX(J)=0.0D0
      DO 6 M=1,NV
    6 BX(J)=BX(J)+XM1(I,M)*B(M+(J-1)*NV)
    8 CONTINUE
      TEMP=0.0D0
      DO 12 M=1,NF
      TEMP=TEMP+XM1(I,NV+M)*B(M+NSTAT*NV)
   12 CONTINUE
      DO 14 J=1,NSTAT
      BX(J)=BX(J)+TEMP
   14 CONTINUE
      DO 20 J=1,NSTAT
      CALL AEXP(BX(J),TEMP)
      P=P+Z(I,J)*(OBS(I)*BX(J)-TIME(I)*TEMP)
   20 CONTINUE
  100 CONTINUE
      P=-P
      RETURN
      END

      SUBROUTINE GRAD1(N,B,G)
C*****************************************************************
C This subroutine computes the first derivative of Q2 (see eqn 2.22).
C Data input: N = dimension of vector B;
C          B = alpha vector;
C    output: G = the derivative of Q2 at B.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF
      INTEGER N
      DIMENSION G(30),B(N),BX(5)
      DO 5 I=1,N
    5 G(I)=0.0D0
      DO 100 I=1,NOBS
      DO 20 J=1,NSTAT
      BX(J)=0.0D0
      DO 10 M=1,NV
   10 BX(J)=BX(J)+XM1(I,M)*B(M+(J-1)*NV)
   20 CONTINUE
      TEMP=0.0D0
      DO 22 M=1,NF
      TEMP=TEMP+XM1(I,M+NV)*B(M+NSTAT*NV)
   22 CONTINUE
      DO 24 J=1,NSTAT
      BX(J)=BX(J)+TEMP
   24 CONTINUE
      DO 32 J=1,NSTAT
      CALL AEXP(BX(J),TEMP)
      DO 30 M=1,NV
      G(M+(J-1)*NV)=G(M+(J-1)*NV)
     C            +Z(I,J)*XM1(I,M)*(OBS(I)-TIME(I)*TEMP)
   30 CONTINUE
   32 CONTINUE
      DO 42 M=1,NF
      DO 40 J=1,NSTAT
      CALL AEXP(BX(J),TEMP)
      G(M+NSTAT*NV)=G(M+NSTAT*NV)
     C            +Z(I,J)*XM1(I,M)*(OBS(I)-TIME(I)*TEMP)
   40 CONTINUE
   42 CONTINUE
  100 CONTINUE
      DO 200 I=1,N
  200 G(I)=-G(I)
      RETURN
      END


      SUBROUTINE MSTEP1(N,B,H,P0,IH,NEVALS,IFAIL,MON)
C*****************************************************************
C  This subroutine is a quasi-Newton algorithm (Nash, 1990) which
C  maximizes the function Q2.
C  Data input: N = dimension of vector B; B = alpha vector;
C          IH = dimension of the corresponding Hessian matrix;
C          NEVALS = # of evaluations for the function Q2;
C    output: H = the Hessian matrix; P0 = maximum value;
```

```
C           B = optimal values of alpha vector.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION B(N), H(IH,N)
      DIMENSION X(30), C(30), G(30), T(30)
      DOUBLE PRECISION K
      INTEGER COUNT
      DATA W,TOL/0.2,1.0D0D-4/,EPS/1.0D0D-6/
      IF (N.LT.0.OR.N.GT.23) GO TO 160
      IFN = N+1
      IG = 1
      RLIM=7.2D0*(10.0D0**74.0D0)
      CALL FUNCT1(N,B,P0)
      IF(P0.GT.RLIM)GOTO180
      CALL GRAD1(N,B,G)
C
C     RESET HESSIAN
C
   10 DO 30 I = 1,N
        DO 20 J = 1,N
   20     H(I,J) = 0.0D0
   30   H(I,I) = 1.0D0D0
      ILAST = IG
C
C     TOP OF ITERATION
C
   40 DO 50 I = 1,N
        X(I) = B(I)
   50   C(I) = G(I)
C
C     FIND SEARCH DIRECTION T
C
      D1 = 0.0D0
      SN=0.0D0
      DO 70 I = 1,N
        S = 0.0D0
        DO 60 J = 1,N
   60     S = S-H(I,J)*G(J)
        T(I) = S
        SN = SN+S*S
   70   D1 = D1-S*G(I)
C
C     CHECK IF DOWNHILL
C
      IF (D1.LE.0.0D0) GO TO 10
C
C     SEARCH ALONG T
C
      SN = 0.5D0D0/DSQRT(SN)
      K = DMIN1(1.0D0D0,SN)
   80 COUNT = 0
      DO 90 I = 1,N
        B(I) = X(I)+K*T(I)
        IF (DABS(B(I)-X(I)).LT.EPS) COUNT = COUNT+1
   90   CONTINUE
C
C     CHECK IF CONVERGED
C
      IF (COUNT.EQ.N) GO TO 150
      CALL FUNCT1(N,B,P)
      IFN = IFN+1
      IF (IFN.GE.NEVALS) GO TO 170
      IF (P.LT.P0-D1*K*TOL) GO TO 100
      K = W*K
      GO TO 80
C
C     NEW LOWEST VALUE
C
  100 P0 = P
      IG = IG+1
      CALL GRAD1(N,B,G)
      IFN = IFN+N
C
C     UPDATE HESSIAN
C
      D1 = 0.0D0
      DO 110 I = 1,N
        T(I) = K*T(I)
```

```
      C(I) = G(I)-C(I)
 110   D1 = D1+T(I)*C(I)
C
C     CHECK IF +VE DEF ADDITION
C
      IF (D1.LE.0.0D0) GO TO 10
      D2 = 0.0D0
      DO 130 I = 1,N
       S = 0.0D0
       DO 120 J = 1,N
 120     S = S+H(I,J)*C(J)
       X(I) = S
 130   D2 = D2+S*C(I)
      D2 = 1+D2/D1
      DO 140 I = 1,N
       DO 140 J = 1,N
 140   H(I,J) = H(I,J)-(T(I)*X(J)+T(J)*X(I)-D2*T(I)*T(J))/D1
      GO TO 40
 150 IFAIL = 0
C SUCCESSFUL CONCLUSION
      RETURN
 160 IFAIL = 1
C N OUT OF RANGE
      RETURN
 170 IFAIL = 2
C TOO MANY FUNCTION EVALUATIONS
      RETURN
 180 IFAIL = 3
C INITIAL POINT INFEASIBLE
      RETURN
 2005 FORMAT( 2X,3G16.4)
      END


      SUBROUTINE ESTEP(NTEMP1,NTEMP2,B,B1)
C*****************************************************************
C  This subroutine executes the E-step of the EM algorithm.
C  Data input: NTEMP1 = dimension of vector B;
C              NTEMP2 = dimension of vector B1;
C              B = beta vector;
C              B1 = alpha vector.
C      Ouput: updated posterior probabilities, Z(I,J).
C*****************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION B(NTEMP1),B1(NTEMP2),TEMP(5),BX(5),BX1(5),TEMPL(5)
      INTEGER NTEMP1,NTEMP2
      SMALL=-79.9D0
      SMALL0=10000000000.0D0
      SMALL0=1.0D0D0/SMALL0
      DO 100 I=1,NOBS
      DO 12 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
      BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
 10 CONTINUE
 12 CONTINUE
      BX(NSTAT)=0.0D0
      DO 18 J=1,NSTAT
      BX1(J)=0.0D0
      DO 16 M=1,NV
      BX1(J)=BX1(J)+XM1(I,M)*B1(M+(J-1)*NV)
 16  CONTINUE
 18  CONTINUE
      IF (NF.EQ.0) GO TO 25
      TEMPP=0.0D0
      DO 22 M=1,NF
      TEMPP=TEMPP+XM1(I,NV+M)*B1(M+NSTAT*NV)
 22  CONTINUE
      DO 24 J=1,NSTAT
      BX1(J)=BX1(J)+TEMPP
 24  CONTINUE
 25  CONTINUE
      CALL AEXP(BX1(1),TEMP(1))
      TEMP1 =BX(1)+OBS(I)*BX1(1)-TIME(I)*TEMP(1)
      DO 30 J=2,NSTAT
      CALL AEXP(BX1(J),TEMP(J))
      TEMP12=(TEMP(J)-TEMP(J-1))*TIME(I)
      TEMP12=(BX(J)-BX(J-1))+OBS(I)*(BX1(J)-BX1(J-1))-TEMP12
```

```
      IF (TEMP12.GT.0.0D0) TEMP1=BX(J)+OBS(I)*BX1(J)-TIME(I)*TEMP(J)
   30 CONTINUE
      TEMP2=0.0D0
      DO 40 J=1,NSTAT
      TEMP(J)=BX(J)+OBS(I)*BX1(J)-TIME(I)*TEMP(J)
      CALL AEXP((TEMP(J)-TEMP1),TEMPL(J))
      TEMP2=TEMP2+TEMPL(J)
   40 CONTINUE
      DO 50 J=1,NSTAT
      Z(I,J)=TEMPL(J)/TEMP2
      IF (Z(I,J).LT.SMALL0) Z(I,J)=0.0D0
   50 CONTINUE
  100 CONTINUE
      RETURN
      END


      SUBROUTINE LLIKELY(NT,BT,F)
C*****************************************************************
C This subroutine computes the observed log likelihood value.
C Data input:  NT = total dimension of vector BT;
C              BT = vector combining beta and alpha vectors.
C     Output:  F = the observed log likelihood value at BT.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      INTEGER NT
      DIMENSION B(30),B1(30),BT(NT),BX(5),BX1(5),TEMP(5)
      DO 1 J=1,N1
    1 B(J)=BT(J)
      DO 2 J=1,N2
    2 B1(J)=BT(N1+J)
      F=0.0D0
      DO 100 I=1,NOBS
      DO 12 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
   10 BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
   12 CONTINUE
      BX(NSTAT)=0.0D0
      DO 18 J=1,NSTAT
      BX1(J)=0.0D0
      DO 16 M=1,NV
      BX1(J)=BX1(J)+XM1(I,M)*B1(M+(J-1)*NV)
   16 CONTINUE
   18 CONTINUE
      IF (NF.EQ.0) GO TO 25
      TEMPP=0.0D0
      DO 22 M=1,NF
      TEMPP=TEMPP+XM1(I,NV+M)*B1(M+NSTAT*NV)
   22 CONTINUE
      DO 24 J=1,NSTAT
      BX1(J)=BX1(J)+TEMPP
   24 CONTINUE
   25 CONTINUE
      CALL AEXP(BX1(1),TEMP(1))
      TEMP1=BX(1)+OBS(I)*BX1(1)-TIME(I)*TEMP(1)
      TEMPP1=BX(1)
      DO 30 J=2,NSTAT
      IF (BX(J).GT.BX(J-1)) TEMPP1=BX(J)
      CALL AEXP(BX1(J),TEMP(J))
      TEMP12=(TEMP(J)-TEMP(J-1))*TIME(I)
      TEMP12=(BX(J)-BX(J-1))+OBS(I)*(BX1(J)-BX1(J-1))-TEMP12
      IF (TEMP12.GT.0.0D0) TEMP1=BX(J)+OBS(I)*BX1(J)-TIME(I)*TEMP(J)
   30 CONTINUE
      TEMP2=0.0D0
      TEMPP2=0.0D0
      DO 40 J=1,NSTAT
      CALL AEXP((BX(J)-TEMPP1),TEMPP21)
      TEMPP2=TEMPP2+TEMPP21
      TEMP12=BX(J)+OBS(I)*BX1(J)-TIME(I)*TEMP(J)
      TEMP12=TEMP12-TEMP1
      CALL AEXP(TEMP12,TEMP22)
      TEMP2=TEMP2+TEMP22
   40 CONTINUE
      F=F+TEMP1+DLOG(TEMP2)-TEMPP1-DLOG(TEMPP2)
  100 CONTINUE
      F=-F
      RETURN
```

```
      END

      SUBROUTINE NEWTON(N,B,H,P0,IH,NEVALS,IFAIL,MON,std)
C*****************************************************************************
C This subroutine is a quasi-Newton algorithm (Nash, 1990) which
C maximizes the observed log likelihood function.
C Data input: N = dimension of vector B;
C             B = vector combining beta and alpha vectors;
C             IH = dimension of the corresponding Hessian matrix;
C             NEVALS = # of evaluations for the observed log likelihood function ;
C     output: H = the Hessian matrix; P0 = maximum value;
C             B = optimal values of alpha vector
C             std = approximate standard errors .
C*****************************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION B(N), H(IH,N),std(30,30)
      DIMENSION X(30), C(30), G(30), T(30)
      DOUBLE PRECISION K
      INTEGER COUNT,ih,n
      DATA W,TOL/0.2,1.0D0D-4/,EPS/1.0D0D-6/
      IF (N.LT.0.OR.N.GT.23) GO TO 160
      IFN = N+1
      IG = 1
      RLIM=7.2D0*(10.0D0**74.0D0)
      CALL LLIKELY(N,B,P0)
      IF (P0.GT.RLIM) GOTO 180
      CALL GLIKELY(N,B,G)
C
C     RESET HESSIAN
C
   10 DO 30 I = 1,N
        DO 20 J = 1,N
   20     H(I,J) = 0.0D0
   30   H(I,I) = 1.0D0D0
      ILAST = IG
C
C     TOP OF ITERATION
C
   40 DO 50 I = 1,N
        X(I) = B(I)
   50   C(I) = G(I)
C
C     FIND SEARCH DIRECTION T
C
      D1 = 0.0D0
      SN=0.0D0
      DO 70 I = 1,N
        S = 0.0D0
        DO 60 J = 1,N
   60     S = S-H(I,J)*G(J)
        T(I) = S
        SN = SN+S*S
   70   D1 = D1-S*G(I)
C
C     CHECK IF DOWNHILL
C
      IF (D1.LE.0.0D0) GO TO 10
C
C     SEARCH ALONG T
C
      SN = 0.5D0D0/DSQRT(SN)
      K = DMIN1(1.0D0D0,SN)
   80 COUNT = 0
      DO 90 I = 1,N
        B(I) = X(I)+K*T(I)
        IF (DABS(B(I)-X(I)).LT.EPS) COUNT = COUNT+1
   90   CONTINUE
C
C     CHECK IF CONVERGED
C
      IF (COUNT.EQ.N) GO TO 150
      CALL LLIKELY(N,B,P)
      IFN = IFN+1
      IF (IFN.GE.NEVALS) GO TO 170
      IF (P.LT.P0-D1*K*TOL) GO TO 100
      K = W*K
      GO TO 80
```

```
C
C     NEW LOWEST VALUE
C
 100 P0 = P
     IG = IG+1
     CALL GLIKELY(N,B,G)
     IFN = IFN+N
C
C     UPDATE HESSIAN
C
     D1 = 0.0D0
     DO 110 I = 1,N
       T(I) = K*T(I)
       C(I) = G(I)-C(I)
 110   D1 = D1+T(I)*C(I)
C
C     CHECK IF +VE DEF ADDITION
C
     IF (D1.LE.0.0D0) GO TO 10
     D2 = 0.0D0
     DO 130 I = 1,N
       S = 0.0D0
       DO 120 J = 1,N
 120     S = S+H(I,J)*C(J)
       X(I) = S
 130   D2 = D2+S*C(I)
     D2 = 1+D2/D1
     DO 140 I = 1,N
       DO 140 J = 1,N
 140   H(I,J) = H(I,J)-(T(I)*X(J)+T(J)*X(I)-D2*T(I)*T(J))/D1
     GO TO 40
 150     do 141 i=1,n
       do 141 j=1,n
 141   std(i,j)=h(i,j)
       IFAIL = 0
C SUCCESSFUL CONCLUSION
     RETURN
 160 IFAIL = 1
C N OUT OF RANGE
     RETURN
 170 IFAIL = 2
C TOO MANY FUNCTION EVALUATIONS
     RETURN
 180 IFAIL = 3
C INITIAL POINT INFEASIBLE
     RETURN
2005 FORMAT( 2X,3G16.4)
     END

     SUBROUTINE GLIKELY(N,B,G)
C*****************************************************************
C This subroutine computes the first derivative of the observed log
C likelihood function.
C Data input: N = dimension of vector B;
C             B = vector combining beta and alpha vectors;
C     Output: G = the derivative of the function at B.
C*****************************************************************
     IMPLICIT DOUBLE PRECISION (A-H,O-Z)
     INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
     DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
     COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
     INTEGER N
     DIMENSION G(30),B(N),BX(5),BX1(5),BTEMP(30),ATEMP(30)
     DO 1 J=1,N1
  1  BTEMP(J)=B(J)
     DO 2 J=1,N2
  2  ATEMP(J)=B(N1+J)
     CALL ESTEP(N1,N2,BTEMP,ATEMP)
     DO 5 I=1,N
  5  G(I)=0.0D0
     DO 100 I=1,NOBS
     DO 12 J=1,NSTAT-1
     BX(J)=0.0D0
     DO 10 M=1,NX
 10  BX(J)=BX(J)+XM(I,M)*BTEMP(M+(J-1)*NX)
 12  CONTINUE
     BX(NSTAT)=0.0D0
     TPROB=0.0D0
     DO 13 J=1,NSTAT
     CALL AEXP(BX(J),TEMP1)
     TPROB=TPROB+TEMP1
```

```
      BX(J)=TEMP1
   13 CONTINUE
      DO 14 J=1,NSTAT
   14 BX(J)=BX(J)/TPROB
      DO 18 J=1,NSTAT
      BX1(J)=0.0D0
      DO 16 M=1,NV
      BX1(J)=BX1(J)+XM1(I,M)*ATEMP(M+(J-1)*NV)
   16 CONTINUE
   18 CONTINUE
      IF (NF.EQ.0) GO TO 22
      TEMPP=0.0D0
      DO 20 M=1,NF
      TEMPP=TEMPP+XM1(I,NV+M)*ATEMP(M+NSTAT*NV)
   20 CONTINUE
      DO 21 J=1,NSTAT
      BX1(J)=BX1(J)+TEMPP
   21 CONTINUE
   22 CONTINUE
      DO 24 J=1,NSTAT-1
      DO 23 M=1,NX
      G(M+(J-1)*NX)=G(M+(J-1)*NX)+XM(I,M)*(Z(I,J)-BX(J))
   23 CONTINUE
   24 CONTINUE
      DO 30 J=1,NSTAT
      CALL AEXP(BX1(J),TRATE)
      BX1(J)=TRATE
   30 CONTINUE
      DO 45 J=1,NSTAT
      DO 42 M=1,NV
      G(M+(J-1)*NV+N1)=G(M+(J-1)*NV+N1)
     C     +XM1(I,M)*Z(I,J)*(OBS(I)-BX1(J))
   42 CONTINUE
   45 CONTINUE
      IF (NF.EQ.0) GO TO 60
      DO 55 M=1,NF
      DO 52 J=1,NSTAT
      G(M+N1+NSTAT*NV)=G(M+N1+NSTAT*NV)
     C     +XM1(I,M+NV)*Z(I,J)*(OBS(I)-BX1(J))
   52 CONTINUE
   55 CONTINUE
   60 CONTINUE
  100 CONTINUE
      DO 200 I=1,N
  200 G(I)=-G(I)
      RETURN
      END


      SUBROUTINE GFIT(NTEMP1,NTEMP2,B,B1,XSQR,FIT,RES,PA,PB,PC)
C*****************************************************************
C  This subroutine computes Pearson statistic, fitted values, Pearson
C  residuals, overdispersion test statistics for each component.
C  Data input: NTEMP1 = dimension of vector B;
C              NTEMP2 = dimension of vector B1;
C              B = beta vector; B1 = alpha vector;
C     Output: XSQR = Pearson statistic;
C              FIT = fitted values including for each component;
C              RES = Pearson residuals including for each component;
C              PA, PB and PC are vectors containing types A, B and C
C              overdisperion test statistics for each component.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      INTEGER NTEMP1,NTEMP2,NCOUNT(10)
      DIMENSION B(NTEMP1),B1(NTEMP2),BX(5),BX1(5),FIT(1000,12)
      DIMENSION PA(10),TPA(10,2)
      DIMENSION CFIT(10),RES(1000,6),PB(10),TPB(10,2),PC(10),TPC(10,2)
      DO 2 J=1,NSTAT
      NCOUNT(J)=0
      TPA(J,1)=0.0D0
      TPA(J,2)=0.0D0
      TPB(J,1)=0.0D0
      TPB(J,2)=0.0D0
      TPC(J,1)=0.0D0
    2 CONTINUE
      XSQR=0.0
      DO 100 I=1,NOBS
      FIT(I,1)=OBS(I)
```

```
      FIT(I,2)=0.0D0
      DO 20 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
   10 BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
   20 CONTINUE
      BX(NSTAT)=0.0D0
      TEMP1=BX(1)
      DO 21 J=2,NSTAT
      IF (BX(J).GT.BX(J-1)) TEMP1=BX(J)
   21 CONTINUE
      TEMPD=0.0D0
      DO 25 J=1,NSTAT
      CALL AEXP((BX(J)-TEMP1),TEMP11)
      BX(J)=TEMP11
      TEMPD=TEMPD+TEMP11
      BX1(J)=0.0D0
      DO 22 M=1,NV
      BX1(J)=BX1(J)+XM1(I,M)*B1(M+(J-1)*NV)
   22 CONTINUE
      if (nv.eq.0) go to 25
      DO 23 M=1,NF
      BX1(J)=BX1(J)+XM1(I,M+NV)*B1(M+NSTAT*NV)
   23 CONTINUE
   24 CALL AEXP(BX1(J),TEMP2)
      BX1(J)=TEMP2*TIME(I)
      FIT(I,2+J)=BX1(J)
      CFIT(J)=BX1(J)
   25 CONTINUE
      DO 26 J=1,NSTAT
      FIT(I,2+NSTAT+J)=BX(J)/TEMPD
      FIT(I,2)=FIT(I,2)+FIT(I,2+J)*FIT(I,2+NSTAT+J)
      RES(I,1+J)=(FIT(I,1)-FIT(I,2+J))*(FIT(I,2+J)**(-0.5D0D0))
   26 CONTINUE
      E2=0.0D0
      DO 30 J=1,NSTAT
      E2=E2+FIT(I,2+NSTAT+J)*(FIT(I,2+J)**2.0D0D0)
   30 CONTINUE
      E2=FIT(I,2)+E2-(FIT(I,2)**(2.0D0D0))
      RES(I,1)=(FIT(I,1)-FIT(I,2))*(E2**(-0.5D0D0))
      XSQR=XSQR+(RES(I,1)**(2.0D0D0))
      NUM=1
      DO 40 J=1,NSTAT-1
      IF (Z(I,J).GT.Z(I,J+1)) GO TO 40
      NUM=J+1
   40 CONTINUE
      NCOUNT(NUM)=NCOUNT(NUM)+1
      TPA(NUM,1)=TPA(NUM,1)+(OBS(I)-CFIT(NUM))**2.0D0-CFIT(NUM)
      TPB(NUM,1)=TPB(NUM,1)+(OBS(I)-CFIT(NUM))**2.0D0-OBS(I)
      TPA(NUM,2)=TPA(NUM,2)+CFIT(NUM)**2.0D0
      TPB(NUM,2)=TPA(NUM,2)
      TPC(NUM,1)=TPC(NUM,1)+((OBS(I)-CFIT(NUM))**2.0D0
     C     -OBS(I))/CFIT(NUM)
  100 CONTINUE
      DO 150 J=1,NSTAT
      PA(J)=TPA(J,1)/((2.0D0D0*TPA(J,2))**0.5D0D0)
      PB(J)=TPB(J,1)/((2.0D0D0*TPB(J,2))**0.5D0D0)
      PC(J)=TPC(J,1)/((DFLOAT(NCOUNT(J))*2.0D0D0))**0.5D0D0
  150 CONTINUE
      RETURN
      END

      SUBROUTINE FLIKELY(NT,BT,F,DRES)
C*********************************************************************
C  This subroutine computes the deviance residuals.
C  Data input: NT = dimension of vector BT;
C              BT = vector combining beta and alpha vectors;
C     Output: DRES = deviance residuals;
C              F = the observed log likelihood function value at BT.
C*********************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NV,NF,N1,N2
      INTEGER NT
      DIMENSION B(30),B1(30),BT(NT),BX(5),BX1(5),TEMP(5),DRES(1000)
      DO 1 J=1,N1
    1 B(J)=BT(J)
      DO 2 J=1,N2
    2 B1(J)=BT(N1+J)
      F=0.0D0
```

```
      DO 100 I=1,NOBS
      DO 12 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
10    BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
12    CONTINUE
      BX(NSTAT)=0.0D0
      DO 18 J=1,NSTAT
      BX1(J)=0.0D0
      DO 16 M=1,NV
      BX1(J)=BX1(J)+XM1(I,M)*B1(M+(J-1)*NV)
16    CONTINUE
18    CONTINUE
      IF (NF.EQ.0) GO TO 25
      TEMPP=0.0D0
      DO 22 M=1,NF
      TEMPP=TEMPP+XM1(I,NV+M)*B1(M+NSTAT*NV)
22    CONTINUE
      DO 24 J=1,NSTAT
      BX1(J)=BX1(J)+TEMPP
24    CONTINUE
25    CONTINUE
      CALL AEXP(BX1(1),TEMP(1))
      TEMP1=BX(1)+OBS(I)*BX1(1)-TIME(I)*TEMP(1)
      TEMPP1=BX(1)
      DO 30 J=2,NSTAT
      IF (BX(J).GT.BX(J-1)) TEMPP1=BX(J)
      CALL AEXP(BX1(J),TEMP(J))
      TEMP12=(TEMP(J)-TEMP(J-1))*TIME(I)
      TEMP12=(BX(J)-BX(J-1))+OBS(I)*(BX1(J)-BX1(J-1))-TEMP12
      IF (TEMP12.GT.0.0D0) TEMP1=BX(J)+OBS(I)*BX1(J)-TIME(I)*TEMP(J)
30    CONTINUE
      TEMP2=0.0D0
      TEMPP2=0.0D0
      DO 40 J=1,NSTAT
      CALL AEXP((BX(J)-TEMPP1),TEMPP21)
      TEMPP2=TEMPP2+TEMPP21
      TEMP12=BX(J)+OBS(I)*BX1(J)-TIME(I)*TEMP(J)
      TEMP12=TEMP12-TEMP1
      CALL AEXP(TEMP12,TEMP22)
      TEMP2=TEMP2+TEMP22
40    CONTINUE
      DRES(I)=TEMP1+DLOG(TEMP2)-TEMPP1-DLOG(TEMPP2)
      F=F+DRES(I)
100   CONTINUE
      F=-F
      RETURN
      END
```

2. Fortran program for computing the maximum likelihood estimates of the mixed logistic regression model.

```
C*********************************************************************
      PROGRAM BINMIX
C *********************************************************************
C  This code find maximum likelihood estimates of the parameters for mixed *
C  binomial regression model. Observed data should be associated with n(i) *
C  which is the number of total trials related to observation i.        *
C  In this code we allow to choose common regression coefficients for    *
C  different components. NVAR = # of different coeffecients              *
C                    NCOM = # of common coeffecients                     *
C  If NCOM = 0, this is the most general case. Note that this code does  *
C  not impose any restriction on mixing probabilities.                   *
C  The program gives the estimated standard errors from the quasi-Newton *
C  approach.                                                             *
C*********************************************************************
C
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION BGUESS(30),OB(30),H(30,30),AGUESS(30),OA(30)
      DIMENSION FIT(1000,13),RES(1000,8),TB(30),TDRES(1000),DRES(1000)
      INTEGER N1,N2,IH1,IH2,MON,NEVALS,IFAIL,NSTEP,NSTEP1,N,IH3
      integer ntol,iter
      dimension tobs(1000),ttime(1000),txm(1000,8),
     c     txm1(1000,8),resl(1000),sign(1000)
      dimension opar(30),se(30),w(1000)
      OPEN( UNIT=1,FILE='lout1')
      OPEN( UNIT=2,FILE='residual')
```

```
      open( unit=3,file='likeres')
      open( unit=8,file='fitout')
      OPEN( UNIT=7,FILE='result')
      OPEN( UNIT=9,FILE='idataout')
      READ(1,100) NOBS,NSTAT,NX,NX1,NVAR,NCOM
       write(*,100) nobs,nstat,nx,nx1,NVAR,NCOM
100   FORMAT(6I5)
      N1=(NSTAT-1)*NX
      N2=NSTAT*NVAR+NCOM
      N=N1+N2
      READ(1,113) (OBS(I),TIME(I), I=1,NOBS)
110   FORMAT(F10.5D0)
       write(*,113) (obs(i),time(i),i=1,nobs)
      do 111 i=1,nobs
      READ (1,112) (XM(I,J),J=1,NX)
       write(*,112) (xm(i,j),j=1,nx)
111   continue
112   FORMAT(6G16.8)
113   FORMAT(2F10.5D0)
      READ(1,110) (BGUESS(I),I=1,N1)
       write(9,110) (bguess(i),i=1,n1)
      DO 115 I=1,NOBS
      READ(1,112) (XM1(I,J),J=1,NX1)
       write (*,112) (xm1(i,j),j=1,nx1)
115   CONTINUE
      READ(1,110) (AGUESS(I),I=1,N2)
       WRITE(9,110) (AGUESS(I),I=1,N2)
      do 118 i=1,nobs
      tobs(i)=obs(i)
      ttime(i)=time(i)
      do 116 j=1,nx
116   txm(i,j)=xm(i,j)
      do 117 j=1,nx1
117   txm1(i,j)=xm1(i,j)
118   continue
      NEVALS=1000
      IH1=N1
      IH2=N2
      IH3=N
      MON=1
      TOL=0.0D01
      toll=0.0D01
      OBSINF=0.0D0
      DEV=0.0D0
      DO 120 I=1,NOBS
120   TDRES(I)=0.0D0
      do 121 i=1,nobs
121   w(i)=0.0D0
      DO 150 I=1,NOBS
      IF (OBS(I).EQ.0.0D0) GO TO 150
      IF (OBS(I).EQ.TIME(I)) GO TO 150
      TDRES(I)=OBS(I)*DLOG(OBS(I))-TIME(I)*DLOG(TIME(I))
     C        +(TIME(I)-OBS(I))*DLOG(TIME(I)-OBS(I))
      DEV=DEV+TDRES(I)
      TEMPSUM=0.0D0
      NSTEP=INT(OBS(I))
      NSTEP1=INT(TIME(I))
      DO 142 J=NSTEP+1,NSTEP1
      TEMPSUM=TEMPSUM+DLOG(DFLOAT(J))
142   CONTINUE
      OBSINF=OBSINF+TEMPSUM
      TEMPSUM=0.0D0
      NSTEP=INT(TIME(I)-OBS(I))
      DO 144 J=1,NSTEP
      TEMPSUM=TEMPSUM+DLOG(DFLOAT(J))
144   CONTINUE
      OBSINF=OBSINF+TEMPSUM
150   CONTINUE
      ntol=nobs
      odev=dev
      do 888 iter=0,ntol
      prel=-(10.0D0**10.0D0)
200   DO 202 I=1,N1
202   OB(I)=BGUESS(I)
      DO 205 I=1,N2
205   OA(I)=AGUESS(I)
      CALL ESTEP(N1,N2,BGUESS,AGUESS)
      CALL MSTEP1(N2,AGUESS,H,P,IH2,NEVALS,IFAIL,MON)
      CALL MSTEP2(N1,BGUESS,H,P,IH1,NEVALS,IFAIL,MON)
      SSR1=0.0D0
      SSR2=0.0D0
```

```
      DO 210 I=1,N1
      SSR1=SSR1+(OB(I)-BGUESS(I))**2.0D0
210   CONTINUE
      DO 220 I=1,N2
220   SSR2=(OA(I)-AGUESS(I))**2.0D0+SSR2
1111  format(4x,2G16.8)
      do 230 i=1,n1
230   tb(i)=bguess(i)
      do 240 i=1,n2
240   tb(i+n1)=aguess(i)
      call llikely(n,tb,f)
      temp=-f-prel
      if (temp.lt.toll) go to 368
      prel=-f
      if (iter.gt.0) go to 363
      f0=-f
      f=-f-obsinf
      write(9,1111) f,f0
      write(*,1111) f,f0
363   IF ((SSR1.GT.TOL).OR.(SSR2.GT.TOL)) GO TO 200
368   call qnewton(n,tb,h,f,nevals,ifail,mon)
      if (iter.eq.0) call flikely(n,tb,f,dres)
      if (iter.gt.0) call llikely(n,tb,f)
      do 375 i=1,n1
375   bguess(i)=tb(i)
      do 376 i=1,n2
376   aguess(i)=tb(i+n1)
      DEV=2*(DEV-(-F))
      if (iter.eq.0) tdev=dev
      if (iter.gt.0) go to 600
      f0=-f
      f=-f-obsinf
      write(9,1111) f,f0
      call estep(n1,n2,bguess,aguess)
      call gfit(n1,n2,bguess,aguess,XSQR,fit,RES)
      temp=dfloat(n1+n2)
      do 378 k=1,n1
      opar(k)=bguess(k)
      se(k)=(h(k,k)**(0.5D0))*temp
378   continue
      do 379 k=1,n2
      opar(k+n1)=aguess(k)
      se(k+n1)=(h(k+n1,k+n1)**(0.5D0))*temp
379   continue
      write(9,4444)
4444  format(4x,'goodness of fit--XSQR, and deviance')
      write(9,7777) XSQR, DEV
      do 380 i=1,nobs
C     IF (OBS(I).EQ.0.0D0) GO TO 380
c     IF (OBS(I).EQ.TIME(I)) GO TO 380
      TEMP=tobs(i)-FIT(I,3)
      IF (TEMP.EQ.0.0D0) sign(I)=0.0D0
      IF (TEMP.NE.0.0D0) sign(i)=TEMP/(ABS(TEMP))
      DRES(I)=(2*(TDRES(I)-DRES(I)))**0.5D0
      DRES(I)=sign(i)*DRES(I)
380   CONTINUE
      do 385 i=1,nobs
      write(8,398) (FIT(I,J),J=1,2+2*NSTAT+1)
      WRITE(2,398) DRES(i), (RES(I,J),J=1,1+NSTAT)
385   continue
398   format(6g18.7)
      do 500 i=1,nobs
      write(7,112) (z(i,j),j=1,nstat)
500   continue
      WRITE(9,5555)
      WRITE(9,7777) (BGUESS(I),(h(i,i)**0.5D0),I=1,N1)
      write(*,7777) (bguess(i),i=1,n1)
      write(9,6666)
      write(9,7777) (AGUESS(I),(h(i+n1,i+n1)**0.5D0),I=1,N2)
      write(*,7777) (aguess(i),i=1,n2)
      go to 603
600   resl(iter)=tdev-dev
      do 601 k=1,n1
      w(iter)=w(iter)+abs(bguess(k)-opar(k))/se(k)
      bguess(k)=opar(k)
601   continue
      do 602 k=1,n2
      w(iter)=w(iter)+abs(aguess(k)-opar(k+n1))/se(k+n1)
      aguess(k)=opar(k+n1)
602   continue
      if (iter.eq.ntol) go to 889
```

```
603  nobs=ntol-1
     dev=odev-tdres(iter+1)
     if (iter.eq.0) go to 614
     do 610 k=1,iter
     obs(k)=tobs(k)
     time(k)=ttime(k)
     do 605 j=1,nx
605  xm(k,j)=txm(k,j)
     do 606 j=1,nx1
606  xm1(k,j)=txm1(k,j)
610  continue
     if (iter.eq.nobs) go to 888
614  do 620 k=iter+1,nobs
     obs(k)=tobs(k+1)
     time(k)=ttime(k+1)
     do 615 j=1,nx
615  xm(k,j)=txm(k+1,j)
     do 616 j=1,nx1
616  xm1(k,j)=txm1(k+1,j)
620  continue
888  continue
889  do 900 i=1,ntol
     resl(i)=sign(i)*(resl(i)**(0.5D0))
     write(3,7778) dres(i),res(i,1),resl(i),w(i)
900  continue
5555 format(4x,'beta-vector')
6666 format(4x,'alpha-vector')
7777 format(4x,2g16.8)
7778 format(4x,4g16.8)
9999 STOP
     END

     SUBROUTINE FUNCT(N,B,P)
C*****************************************************************
C This subroutine computes the value of function Q1 in Chapter 3.
C Data input: N = dimension of vector B;
C            B = beta vector;
C    output: P = the function value Q1(B).
C*****************************************************************
     IMPLICIT DOUBLE PRECISION (A-H,O-Z)
     INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
     DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
     COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
     INTEGER N
     DIMENSION B(N),BX(5)
     P=0.0D0
     DO 100 I=1,NOBS
     DO 8 J=1,NSTAT-1
     BX(J)=0.0D0
     DO 6 M=1,NX
  6  BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
  8  CONTINUE
     BX(NSTAT)=0.0D0
     TEMPMAX=BX(1)
C*********************************************
C  Loop 20 finds the largest BX(J), TEMPMAX. *
C*********************************************
     DO 20 J=2,NSTAT
     IF (BX(J).GT.BX(J-1)) TEMPMAX=BX(J)
 20  CONTINUE
     TEMPSUM=0.0D0
     DO 30 J=1,NSTAT
     P=P+Z(I,J)*BX(J)
     CALL AEXP((BX(J)-TEMPMAX),TEMP3)
     TEMPSUM=TEMPSUM+TEMP3
 30  CONTINUE
     P=P-TEMPMAX-DLOG(TEMPSUM)
100  CONTINUE
     P=-P
     RETURN
     END

     SUBROUTINE GRAD(N,B,G)
C*******************************************************************
C This subroutine computes the first derivative of Q1 (see eqn 3.18).
C Data input: N = dimension of vector B;
C            B = beta vector;
C    output: G = the derivative of Q1 at B.
C*******************************************************************
     IMPLICIT DOUBLE PRECISION (A-H,O-Z)
     INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
```

```
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      INTEGER N
      DIMENSION G(25),B(N),TEMP(25),BX(5)
      DO 5 I=1,N
    5 G(I)=0.0D0
      DO 100 I=1,NOBS
      DO 20 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
   10 BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
   20 CONTINUE
      BX(NSTAT)=0.0D0
      TEMPMAX=BX(1)
C*****************************************
C   Loop 30 finds the largest BX(J), TEMPMAX. *
C*****************************************
      DO 30 J=2,NSTAT
      IF (BX(J).GT.BX(J-1)) TEMPMAX=BX(J)
   30 CONTINUE
      TEMPSUM=0.0D0
      DO 40 J=1,NSTAT
      BX(J)=BX(J)-TEMPMAX
      CALL AEXP(BX(J),TEMP(J))
      TEMPSUM=TEMPSUM+TEMP(J)
   40 CONTINUE
      DO 60 J=1,NSTAT-1
      TEMPPRO=TEMP(J)/TEMPSUM
      DO 50 M=1,NX
      G(M+(J-1)*NX)=G(M+(J-1)*NX)+XM(I,M)*(Z(I,J)-TEMPPRO)
   50 CONTINUE
   60 CONTINUE
  100 CONTINUE
      DO 200 I=1,N
  200 G(I)=-G(I)
      RETURN
      END


      SUBROUTINE MSTEP2(N,B,H,P0,IH,NEVALS,IFAIL,MON)
C*************************************************************
C This subroutine is a quasi-Newton algorithm (Nash, 1990) which
C maximizes the function Q1.
C Data input: N = dimension of vector B; B = beta vector;
C             IH = dimension of the Hessian matrix;
C             NEVALS = # of evaluations for the function Q1;
C     output: H = the Hessian matrix; P0 = maximum value;
C             B = optimal values of beta vector.
C*************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION B(N), H(IH,N)
      DIMENSION X(30), C(30), G(30), T(30)
      DOUBLE PRECISION K
      INTEGER COUNT
      DATA W,TOL/0.2,1.0D0D-4/,EPS/1.0D0D-6/
      IF (N.LT.0.OR.N.GT.23) GO TO 160
      IFN = N+1
      IG = 1
      RLIM=7.2*(10.0D0**74.0)
      CALL FUNCT(N,B,P0)
      IF(P0.GT.RLIM)GOTO180
      CALL GRAD(N,B,G)
C
C     RESET HESSIAN
C
   10 DO 30 I = 1,N
        DO 20 J = 1,N
   20     H(I,J) = 0.0D0
   30   H(I,I) = 1.0D0
      ILAST = IG
C
C     TOP OF ITERATION
C
   40 DO 50 I = 1,N
        X(I) = B(I)
   50   C(I) = G(I)
C
C     FIND SEARCH DIRECTION T
```

```
C
      D1 = 0.0D0
      SN=0.0D0
      DO 70 I = 1,N
        S = 0.0D0
        DO 60 J = 1,N
   60     S = S-H(I,J)*G(J)
        T(I) = S
        SN = SN+S*S
   70   D1 = D1-S*G(I)
C
C     CHECK IF DOWNHILL
C
      IF (D1.LE.0.0D0) GO TO 10
C
C     SEARCH ALONG T
C
      SN = 0.5D0/DSQRT(SN)
      K = DMIN1(1.0D0D0,SN)
   80 COUNT = 0
      DO 90 I = 1,N
        B(I) = X(I)+K*T(I)
        IF (DABS(B(I)-X(I)).LT.EPS) COUNT = COUNT+1
   90   CONTINUE
C
C     CHECK IF CONVERGED
C
      IF (COUNT.EQ.N) GO TO 150
      CALL FUNCT(N,B,P)
      IFN = IFN+1
      IF (IFN.GE.NEVALS) GO TO 170
      IF (P.LT.P0-D1*K*TOL) GO TO 100
      K = W*K
      GO TO 80
C
C     NEW LOWEST VALUE
C
  100 P0 = P
      IG = IG+1
      CALL GRAD(N,B,G)
      IFN = IFN+N
C
C     UPDATE HESSIAN
C
      D1 = 0.0D0
      DO 110 I = 1,N
        T(I) = K*T(I)
        C(I) = G(I)-C(I)
  110   D1 = D1+T(I)*C(I)
C
C     CHECK IF +VE DEF ADDITION
C
      IF (D1.LE.0.0D0D0) GO TO 10
      D2 = 0.0D0
      DO 130 I = 1,N
        S = 0.0D0
        DO 120 J = 1,N
  120     S = S+H(I,J)*C(J)
        X(I) = S
  130   D2 = D2+S*C(I)
      D2 = 1+D2/D1
      DO 140 I = 1,N
        DO 140 J = 1,N
  140   H(I,J) = H(I,J)-(T(I)*X(J)+T(J)*X(I)-D2*T(I)*T(J))/D1
      GO TO 40
  150 IFAIL = 0
C SUCCESSFUL CONCLUSION
      RETURN
  160 IFAIL = 1
C N OUT OF RANGE
      RETURN
  170 IFAIL = 2
C TOO MANY FUNCTION EVALUATIONS
      RETURN
  180 IFAIL = 3
C INITIAL POINT INFEASIBLE
      RETURN
 2005 FORMAT( 2X,3G16.4)
      END
```

```
      SUBROUTINE AEXP(X,F)
C*****************************************************************
C This subroutine computes a exponential function value.
C Data input: X = real number;
C     output: F = exp(X).
C*****************************************************************
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      INTEGER NSTEP
      TEMP1=ABS(X)
      IF (TEMP1.GT.79.9) GO TO 50
      F=DEXP(X)
      GO TO 200
   50 IF (X.LT.-79.9) GO TO 150
      IF (X.GT.150.0D0) X=150.0D0
      F=1.0D0+X
      NSTEP=1
      FACTOR=1.0D0
      TEMP1=DFLOAT(NSTEP)
      TEMP2=X/TEMP1
  100 IF (TEMP2.LT.1.0D0) GO TO 200
      NSTEP=NSTEP+1
      TEMP1=DFLOAT(NSTEP)
      FACTOR=X/TEMP1
      TEMP2=TEMP2*FACTOR
      F=F+TEMP2
      GO TO 100
  150 F=0.0D0
  200 RETURN
      END


      SUBROUTINE FUNCT1(N,B,P)
C*****************************************************************
C This subroutine computes the value of function Q2 in Chapter 3.
C Data input: N = dimension of vector B;
C             B = alpha vector;
C     output: P = the function value Q2(B).
C*****************************************************************
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      INTEGER N
      DIMENSION B(N),BX(5)
      P=0.0D0
      TEMP1=1.0D0
      DO 100 I=1,NOBS
C*****************************************************************
C   Loop 8 computes BX(J) for variable coefficient part. *
C*****************************************************************
      DO 8 J=1,NSTAT
      BX(J)=0.0D0
      DO 6 M=1,NVAR
    6 BX(J)=BX(J)+XM1(I,M)*B(M+(J-1)*NVAR)
    8 CONTINUE
      IF (NCOM.EQ.0) GO TO 11
      DO 10 J=1,NSTAT
      DO 9 M=1,NCOM
      BX(J)=BX(J)+XM1(I,NVAR+M)*B(NVAR*NSTAT+M)
    9 CONTINUE
   10 CONTINUE
   11 CONTINUE
      DO 20 J=1,NSTAT
      IF (BX(J).LT.0.0D0) GO TO 15
      CALL AEXP(-BX(J),TEMP)
      P=P+Z(I,J)*((OBS(I)-TIME(I))*BX(J)-TIME(I)*DLOG(TEMP1+TEMP))
      GO TO 20
   15 CALL AEXP(BX(J),TEMP)
      P=P+Z(I,J)*(OBS(I)*BX(J)-TIME(I)*DLOG(TEMP1+TEMP))
   20 CONTINUE
  100 CONTINUE
      P=-P
      RETURN
      END

      SUBROUTINE GRAD1(N,B,G)
C*****************************************************************
C This subroutine computes the first derivative of Q2 (see eqn 3.19).
C Data input: N = dimension of vector B;
C             B = alpha vector;
```

```
C     output: G = the derivative of Q2 at B.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      INTEGER N
      DIMENSION G(30),B(N),BX(5)
      TEMP1=1.0D0
      DO 5 I=1,N
    5 G(I)=0.0D0
      DO 100 I=1,NOBS
C*****************************************************
C  Loop 20 computes BX(J) for variable coefficient part. *
C*****************************************************
      DO 20 J=1,NSTAT
      BX(J)=0.0D0
      DO 10 M=1,NVAR
   10 BX(J)=BX(J)+XM1(I,M)*B(M+(J-1)*NVAR)
   20 CONTINUE
      IF (NCOM.EQ.0) GO TO 25
      DO 24 J=1,NSTAT
      DO 22 M=1,NCOM
      BX(J)=BX(J)+XM1(I,NVAR+M)*B(NSTAT*NVAR+M)
   22 CONTINUE
   24 CONTINUE
   25 CONTINUE
C*********************************************
C  Loop 40 computes the gradient.          *
C*********************************************
      DO 40 J=1,NSTAT
      IF (BX(J).GT.0.0D0) GO TO 35
      CALL AEXP(BX(J),TEMP)
      DO 30 M=1,NVAR
      G(M+(J-1)*NVAR)=G(M+(J-1)*NVAR)
     C     +Z(I,J)*XM1(I,M)*(OBS(I)-TIME(I)*TEMP/(TEMP1+TEMP))
   30 CONTINUE
      GO TO 40
   35 CALL AEXP(-BX(J),TEMP)
      DO 36 M=1,NX1
      G(M+(J-1)*NVAR)=G(M+(J-1)*NVAR)
     C     +Z(I,J)*XM1(I,M)*(OBS(I)-TIME(I)/(TEMP1+TEMP))
   36 CONTINUE
   40 CONTINUE
      IF (NCOM.EQ.0) GO TO 81
      DO 80 M=1,NCOM
      DO 60 J=1,NSTAT
      IF (BX(J).GT.0.0D0) GO TO 55
      CALL AEXP(BX(J),TEMP)
      G(M+NVAR*NSTAT)=G(M+NVAR*NSTAT)
     C  +Z(I,J)*XM1(I,M+NVAR)*(OBS(I)-TIME(I)*TEMP/(TEMP1+TEMP))
      GO TO 60
   55 CALL AEXP(-BX(J),TEMP)
      G(M+(J-1)*NVAR)=G(M+(J-1)*NVAR)
     C     +Z(I,J)*XM1(I,M+NVAR)*(OBS(I)-TIME(I)/(TEMP1+TEMP))
   60 CONTINUE
   80 CONTINUE
   81 CONTINUE
  100 CONTINUE
      DO 200 I=1,N
  200 G(I)=-G(I)
      RETURN
      END


      SUBROUTINE MSTEP1(N,B,H,P0,IH,NEVALS,IFAIL,MON)
C*****************************************************************
C  This subroutine is a quasi-Newton algorithm (Nash, 1990) which
C  maximizes the function Q2.
C  Data input: N = dimension of vector B; B = alpha vector;
C              IH = dimension of the corresponding Hessian matrix;
C              NEVALS = # of evaluations for the function Q2;
C     output: H = the Hessian matrix; P0 = maximum value;
C              B = optimal values of alpha vector.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION B(N), H(IH,N)
      DIMENSION X(30), C(30), G(30), T(30)
```

```
      DOUBLE PRECISION K
      INTEGER COUNT
      DATA W,TOL/0.2,1.0D0D-4/,EPS/1.0D0D-6/
      IF (N.LT.0.OR.N.GT.23) GO TO 160
      IFN = N+1
      IG = 1
      RLIM=7.2*(10.0D0**74.0)
      CALL FUNCT1(N,B,P0)
      IF(P0.GT.RLIM)GOTO180
      CALL GRAD1(N,B,G)
C
C     RESET HESSIAN
C
   10 DO 30 I = 1,N
        DO 20 J = 1,N
   20     H(I,J) = 0.0D0
   30   H(I,I) = 1.0D0
      ILAST = IG
C
C     TOP OF ITERATION
C
   40 DO 50 I = 1,N
        X(I) = B(I)
   50   C(I) = G(I)
C
C     FIND SEARCH DIRECTION T
C
      D1 = 0.0D0
      SN=0.0D0
      DO 70 I = 1,N
        S = 0.0D0
        DO 60 J = 1,N
   60     S = S-H(I,J)*G(J)
        T(I) = S
        SN = SN+S*S
   70   D1 = D1-S*G(I)
C
C     CHECK IF DOWNHILL
C
      IF (D1.LE.0.0D0) GO TO 10
C
C     SEARCH ALONG T
C
      SN = 0.5D0/DSQRT(SN)
      K = DMIN1(1.0D0D0,SN)
   80 COUNT = 0
      DO 90 I = 1,N
        B(I) = X(I)+K*T(I)
        IF (DABS(B(I)-X(I)).LT.EPS) COUNT = COUNT+1
   90   CONTINUE
C
C     CHECK IF CONVERGED
C
      IF (COUNT.EQ.N) GO TO 150
      CALL FUNCT1(N,B,P)
      IFN = IFN+1
      IF (IFN.GE.NEVALS) GO TO 170
      IF (P.LT.P0-D1*K*TOL) GO TO 100
      K = W*K
      GO TO 80
C
C     NEW LOWEST VALUE
C
  100 P0 = P
      IG = IG+1
      CALL GRAD1(N,B,G)
      IFN = IFN+N
C
C     UPDATE HESSIAN
C
      D1 = 0.0D0
      DO 110 I = 1,N
        T(I) = K*T(I)
        C(I) = G(I)-C(I)
  110   D1 = D1+T(I)*C(I)
C
C     CHECK IF +VE DEF ADDITION
C
      IF (D1.LE.0.0D0D0) GO TO 10
      D2 = 0.0D0
      DO 130 I = 1,N
```

```
      S = 0.0D0
      DO 120 J = 1,N
  120   S = S+H(I,J)*C(J)
      X(I) = S
  130 D2 = D2+S*C(I)
    D2 = 1+D2/D1
    DO 140 I = 1,N
      DO 140 J = 1,N
  140 H(I,J) = H(I,J)-(T(I)*X(J)+T(J)*X(I)-D2*T(I)*T(J))/D1
    GO TO 40
  150 IFAIL = 0
C SUCCESSFUL CONCLUSION
    RETURN
  160 IFAIL = 1
C N OUT OF RANGE
    RETURN
  170 IFAIL = 2
C TOO MANY FUNCTION EVALUATIONS
    RETURN
  180 IFAIL = 3
C INITIAL POINT INFEASIBLE
    RETURN
 2005 FORMAT( 2X,3G16.4)
    END


    SUBROUTINE ESTEP(N1,N2,B,B1)
C*****************************************************************
C  This subroutine executes the E-step of the EM algorithm.
C  Data input: NTEMP1 = dimension of vector B;
C            NTEMP2 = dimension of vector B1;
C            B = beta vector;
C            B1 = alpha vector.
C      Ouput: updated posterior probabilities, Z(I,J).
C*****************************************************************
    IMPLICIT DOUBLE PRECISION(A-H,O-Z)
    INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
    DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
    COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
    INTEGER N1,N2
    DIMENSION B(N1),B1(N2),TEMP(5),BX(5),BX1(5)
    SMALL0=10000000000.0D0
    SMALL0=1.0D0/SMALL0
    ONE=1.0D0
    DO 100 I=1,NOBS
C*****************************
C  Loop 12 computes BX(J).   *
C*****************************
    DO 12 J=1,NSTAT-1
    BX(J)=0.0D0
    DO 10 M=1,NX
    BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
  10 CONTINUE
  12 CONTINUE
    BX(NSTAT)=0.0D0
C*****************************************************
C  Loop 21 computes BX1(J) for variable coefficient part.  *
C*****************************************************
    DO 21 J=1,NSTAT
    BX1(J)=0.0D0
    DO 20 M=1,NVAR
  20 BX1(J)=BX1(J)+XM1(I,M)*B1(M+(J-1)*NVAR)
  21 CONTINUE
    IF (NCOM.EQ.0) GO TO 24
    DO 23 J=1,NSTAT
    DO 22 M=1,NCOM
    BX1(J)=BX1(J)+XM1(I,M+NVAR)*B1(NSTAT*NVAR+M)
  22 CONTINUE
  23 CONTINUE
  24 CONTINUE
C*****************************************************
C  Loop 30 finds the largest item in exponential functions.   *
C*****************************************************
    IF (BX1(1).GT.0.0D0) GO TO 25
    CALL AEXP(BX1(1),TEMP1)
    TEMP(1)=BX(1)+OBS(I)*BX1(1)-TIME(I)*DLOG(ONE+TEMP1)
    GO TO 26
 25 CALL AEXP(-BX1(1),TEMP1)
    TEMP(1)=BX(1)+(OBS(I)-TIME(I))*BX1(1)-TIME(I)*DLOG(ONE+TEMP1)
 26 TEMPMAX=TEMP(1)
    DO 34 J=2,NSTAT
```

```
      IF (BX1(J).GT.0.0D0) GO TO 32
      CALL AEXP(BX1(J),TEMP1)
      TEMP(J)=BX(J)+OBS(I)*BX1(J)-TIME(I)*DLOG(ONE+TEMP1)
      GO TO 33
   32 CALL AEXP(-BX1(J),TEMP1)
      TEMP(J)=BX(J)+(OBS(I)-TIME(I))*BX1(J)-TIME(I)*DLOG(ONE+TEMP1)
   33 IF (TEMP(J).GT.TEMP(J-1)) TEMPMAX=TEMP(J)
   34 CONTINUE
C*********************************************
C  Loops 40 and 50 compute Z(I,J) values.     *
C*********************************************
      TEMPSUM=0.0D0
      DO 40 J=1,NSTAT
      TEMPP=TEMP(J)-TEMPMAX
      CALL AEXP(TEMPP,TEMP(J))
      TEMPSUM=TEMPSUM+TEMP(J)
   40 CONTINUE
      DO 50 J=1,NSTAT
      Z(I,J)=TEMP(J)/TEMPSUM
      IF (Z(I,J).LT.SMALL0) Z(I,J)=0.0D0
   50 CONTINUE
  100 CONTINUE
      RETURN
      END

      SUBROUTINE GFIT(N1,N2,B,B1,XSQR,FIT,RES)
C*****************************************************************
C This subroutine computes Pearson statistic, fitted values, Pearson
C residuals.
C Data input: N1 = dimension of vector B;
C             N2 = dimension of vector B1;
C             B = beta vector; B1 = alpha vector;
C    Output: XSQR = Pearson statistic;
C             FIT = fitted values including for each component;
C             RES = Pearson residuals including for each component;
C*****************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      INTEGER N1,N2
      DIMENSION B(N1),B1(N2),BX(5),BX1(5),TEMP(5)
      DIMENSION FIT(1000,13),RES(1000,8)
      ONE=1.0D0
      XSQR=0.0D0
      DO 100 I=1,NOBS
      FIT(I,1+1)=OBS(I)
      FIT(I,2+1)=0.0D0
C*****************************
C  Loop 12 computes BX(J).    *
C*****************************
      DO 12 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
      BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
   10 CONTINUE
   12 CONTINUE
      BX(NSTAT)=0.0D0
      TEMP1=BX(1)
      DO 14 J=2,NSTAT
      IF (BX(J).GT.BX(J-1)) TEMP1=BX(J)
   14 CONTINUE
      TEMPD=0.0D0
      DO 16 J=1,NSTAT
      CALL AEXP((BX(J)-TEMP1),TEMP11)
      BX(J)=TEMP11
      TEMPD=TEMPD+TEMP11
   16 CONTINUE
      DO 18 J=1,NSTAT
      FIT(I,2+NSTAT+J+1)=BX(J)/TEMPD
   18 CONTINUE
C*****************************************************
C  Loop 21 computes BX1(J) for variable coefficient part. *
C*****************************************************
      DO 21 J=1,NSTAT
      BX1(J)=0.0D0
      DO 20 M=1,NVAR
   20 BX1(J)=BX1(J)+XM1(I,M)*B1(M+(J-1)*NVAR)
   21 CONTINUE
      IF (NCOM.EQ.0) GO TO 24
      DO 23 J=1,NSTAT
```

```
      DO 22 M=1,NCOM
      BX1(J)=BX1(J)+XM1(I,M+NVAR)*B1(NSTAT*NVAR+M)
   22 CONTINUE
   23 CONTINUE
   24 CONTINUE
C*********************************************************
      E1=0.0D0
      DO 40 J=1,NSTAT
      IF (BX1(J).LT.0.0D0) GO TO 35
      CALL AEXP(-BX1(J),TEMPP)
      TEMP(J)=ONE/(ONE+TEMPP)
      TEMP(J)=TIME(I)*TEMP(J)*(ONE-TEMP(J))
      E1=E1+FIT(I,2+NSTAT+J+1)*TEMP(J)
      FIT(I,2+J+1)=TIME(I)/(ONE+TEMPP)
      GO TO 40
   35 CALL AEXP(BX1(J),TEMPP)
      TEMP(J)=TEMPP/(ONE+TEMPP)
      TEMP(J)=TIME(I)*TEMP(J)*(ONE-TEMP(J))
      E1=E1+FIT(I,2+NSTAT+J+1)*TEMP(J)
      FIT(I,2+J+1)=TIME(I)*TEMPP/(ONE+TEMPP)
   40 CONTINUE
      DO 42 J=1,NSTAT
      FIT(I,2+1)=FIT(I,2+1)+FIT(I,2+NSTAT+J+1)*FIT(I,2+J+1)
      RES(I,1+J)=(FIT(I,1+1)-FIT(I,2+J+1))*(TEMP(J)**(-0.5D0))
   42 CONTINUE
      E2=0.0D0
      DO 50 J=1,NSTAT
      E2=E2+FIT(I,2+NSTAT+J+1)*(FIT(I,2+J+1)**(2.0D0))
   50 CONTINUE
      E2=E1+E2-(FIT(I,2+1)**(2.0D0))
      FIT(I,1)=E2
      RES(I,1)=(FIT(I,1+1)-FIT(I,2+1))*(E2**(-0.5D0))
      XSQR=XSQR+(RES(I,1)**(2.0D0))
  100 CONTINUE
      RETURN
      END


      SUBROUTINE GLIKELY(N,TB,G)
C*********************************************************
C This subroutine computes the first derivative of the observed log
C likelihood function.
C Data input: N = dimension of vector B;
C             B = vector combining beta and alpha vectors;
C    Output: G = the derivative of the function at B.
C*********************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),
     C  Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      INTEGER N,N1,N2
      DIMENSION TB(N),B(25),B1(25),BX(5),BX1(5),TEMP(5),COM(5)
     C        ,G(25),P(5)
      DO 1 I=1,N
    1 G(I)=0.0D0
      N1=(NSTAT-1)*NX
      N2=NSTAT*NVAR+NCOM
      DO 2 I=1,N1
    2 B(I)=TB(I)
      DO 3 I=1,N2
    3 B1(I)=TB(N1+I)
      ONE=1.0D0
      DO 100 I=1,NOBS
C*****************************
C  Loop 20 computes BX(J).  *
C*****************************
      DO 20 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
   10 BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
   20 CONTINUE
      BX(NSTAT)=0.0D0
      DO 22 J=1,NSTAT
      BX1(J)=0.0D0
C*********************************************************
C  Loop 21 computes BX1(J) for variable coefficient part.  *
C*********************************************************
      DO 21 M=1,NVAR
      BX1(J)=BX1(J)+XM1(I,M)*B1(M+(J-1)*NVAR)
   21 CONTINUE
```

```
 22 CONTINUE
    IF (NCOM.EQ.0) GO TO 25
    DO 24 J=1,NSTAT
    DO 23 M=1,NCOM
    BX1(J)=BX1(J)+XM1(I,M+NVAR)*B1(M+NSTAT*NVAR)
 23 CONTINUE
 24 CONTINUE
 25 CONTINUE
C*****************************************************
    PMAX=BX(1)
    DO 26 J=2,NSTAT
    IF (BX(J).GT.BX(J-1)) PMAX=BX(J)
 26 CONTINUE
    PSUM=0.0D0
    DO 28 J=1,NSTAT
    bx(j)=bx(j)-pmax
    CALL AEXP(BX(J),P(J))
    PSUM=PSUM+P(J)
 28 CONTINUE
C*****************************************
C CALCULATE MIXING PROBABILITIES P_j     *
C*****************************************
    DO 29 J=1,NSTAT
 29 P(J)=P(J)/PSUM
C*****************************************
C CALCULATE BINOMIAL PARAMETERS THETA_j  *
C*****************************************
    DO 40 J=1,NSTAT
    IF (BX1(J).LT.0.0D0) GO TO 35
    CALL AEXP(-BX1(J),TEMPP)
    TEMP(J)=BX(J)+(OBS(I)-TIME(I))*BX1(J)-TIME(I)*DLOG(ONE+TEMPP)
    BX1(J)=1.0D0/(1.0D0+TEMPP)
    GO TO 40
 35 CALL AEXP(BX1(J),TEMPP)
    TEMP(J)=BX(J)+OBS(I)*BX1(J)-TIME(I)*DLOG(ONE+TEMPP)
    BX1(J)=TEMPP/(1.0D0+TEMPP)
 40 CONTINUE
    TEMPMAX=TEMP(1)
    DO 45 J=2,NSTAT
    IF (TEMP(J).GT.TEMP(J-1)) TEMPMAX=TEMP(J)
 45 CONTINUE
    TEMPSUM=0.0D0
    DO 48 J=1,NSTAT
    TEMPP=TEMP(J)-TEMPMAX
    CALL AEXP(TEMPP,COM(J))
    TEMPSUM=TEMPSUM+COM(J)
 48 CONTINUE
    DO 50 J=1,NSTAT
    COM(J)=COM(J)/TEMPSUM
 50 CONTINUE
    DO 70 J=1,NSTAT-1
    TEMPP=COM(J)-P(J)
    DO 65 M=1,NX
    G(M+NX*(J-1))=G(M+NX*(J-1))+XM(I,M)*TEMPP
 65 CONTINUE
 70 CONTINUE
    TEMPSUM=0.0D0
    DO 80 J=1,NSTAT
    TEMPP=COM(J)*(OBS(I)-TIME(I)*BX1(J))
    TEMPSUM=TEMPSUM+TEMPP
    DO 75 M=1,NVAR
    G(M+N1+NVAR*(J-1))=G(M+N1+NVAR*(J-1))+TEMPP*XM1(I,M)
 75 CONTINUE
 80 CONTINUE
    IF (NCOM.EQ.0) GO TO 100
    DO 90 M=1,NCOM
    G(M+N1+NVAR*NSTAT)=G(M+N1+NVAR*NSTAT)+
   C              TEMPSUM*XM1(I,M+NVAR)
 90 CONTINUE
 100 CONTINUE
    do 200 i=1,n
 200 g(i)=-g(i)
    RETURN
    END


    SUBROUTINE LLIKELY(N,TB,F)
C*****************************************************************
C This subroutine computes the observed log likelihood value.
C Data input:  N = total dimension of vector BT;
C              TB = vector combining beta and alpha vectors.
```

```
C     Output:  F = the observed log likelihood value at BT.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),
     C          Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      INTEGER N1,N2
      DIMENSION TB(N),B(25),B1(25),BX(5),BX1(5),TEMP(5)
      N1=(NSTAT-1)*NX
      N2=NSTAT*NVAR+NCOM
      DO 1 I=1,N1
    1 B(I)=TB(I)
      DO 2 I=1,N2
    2 B1(I)=TB(N1+I)
      F=0.0D0
      ONE=1.0D0
      DO 100 I=1,NOBS
C*****************************
C  Loop 20 computes BX(J).  *
C*****************************
      DO 20 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
   10 BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
   20 CONTINUE
      BX(NSTAT)=0.0D0
      DO 22 J=1,NSTAT
      BX1(J)=0.0D0
C*****************************************************************
C  Loop 21 computes BX1(J) for variable coefficient part.   *
C*****************************************************************
      DO 21 M=1,NVAR
      BX1(J)=BX1(J)+XM1(I,M)*B1(M+(J-1)*NVAR)
   21 CONTINUE
   22 CONTINUE
      IF (NCOM.EQ.0) GO TO 25
      DO 24 J=1,NSTAT
      DO 23 M=1,NCOM
      BX1(J)=BX1(J)+XM1(I,M+NVAR)*B1(M+NSTAT*NVAR)
   23 CONTINUE
   24 CONTINUE
   25 CONTINUE
C*****************************************************
      PMAX=BX(1)
      DO 26 J=2,NSTAT
      IF (BX(J).GT.BX(J-1)) PMAX=BX(J)
   26 CONTINUE
      PSUM=0.0D0
      DO 28 J=1,NSTAT
      bx(j)=bx(j)-pmax
      CALL AEXP(BX(J),TEMPP )
      PSUM=PSUM+TEMPP
   28 CONTINUE
      F=F-DLOG(PSUM)
C*****************************************************
      DO 40 J=1,NSTAT
      IF (BX1(J).LT.0.0D0) GO TO 35
      CALL AEXP(-BX1(J),TEMPP)
      TEMP(J)=BX(J)+(OBS(I)-TIME(I))*BX1(J)-TIME(I)*DLOG(ONE+TEMPP)
      GO TO 40
   35 CALL AEXP(BX1(J),TEMPP)
      TEMP(J)=BX(J)+OBS(I)*BX1(J)-TIME(I)*DLOG(ONE+TEMPP)
   40 CONTINUE
      TEMPMAX=TEMP(1)
      DO 45 J=2,NSTAT
      IF (TEMP(J).GT.TEMP(J-1)) TEMPMAX=TEMP(J)
   45 CONTINUE
      TEMPSUM=0.0D0
      DO 48 J=1,NSTAT
      TEMPP=TEMP(J)-TEMPMAX
      CALL AEXP(TEMPP,TEMP2)
      TEMPSUM=TEMPSUM+TEMP2
   48 CONTINUE
      F=F+TEMPMAX+DLOG(TEMPSUM)
  100 CONTINUE
      f=-f
      RETURN
      END
```

```
      SUBROUTINE QNEWTON(N,B,H,P0,NEVALS,IFAIL,MON)
C*****************************************************************************
C  This subroutine is a quasi-Newton algorithm (Nash, 1990) which
C  maximizes the observed log likelihood function.
C  Data input: N = dimension of vector B;
C            B = vector combining beta and alpha vectors;
C            IH = dimension of the corresponding Hessian matrix;
C            NEVALS = # of evaluations for the observed log likelihood function ;
C     output: H = the Hessian matrix; P0 = maximum value;
C            B = optimal values of alpha vector
C*****************************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),
     C      Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION B(N), H(30,30)
      DIMENSION X(30), C(30), G(30), T(30)
      DOUBLE PRECISION K
      INTEGER COUNT,IH,N
      DATA W,TOL/0.2,1.0D0D-4/,EPS/1.0D0D-6/
      ih=n
      IF (N.LT.0.OR.N.GT.23) GO TO 160
      IFN = N+1
      IG = 1
      RLIM=7.2D0*(10.0D0**74.0D0)
      CALL LLIKELY(N,B,P0)
      IF(P0.GT.RLIM)GOTO180
      CALL GLIKELY(N,B,G)
C
C     RESET HESSIAN
C
   10 DO 30 I = 1,N
        DO 20 J = 1,N
   20     H(I,J) = 0.0D0
   30   H(I,I) = 1.0D0
      ILAST = IG
C
C     TOP OF ITERATION
C
   40 DO 50 I = 1,N
        X(I) = B(I)
   50   C(I) = G(I)
C
C     FIND SEARCH DIRECTION T
C
      D1 = 0.0D0
      SN=0.0D0
      DO 70 I = 1,N
        S = 0.0D0
        DO 60 J = 1,N
   60     S = S-H(I,J)*G(J)
        T(I) = S
        SN = SN+S*S
   70   D1 = D1-S*G(I)
C
C     CHECK IF DOWNHILL
C
      IF (D1.LE.0.0D0) GO TO 10
C
C     SEARCH ALONG T
C
      SN = 0.5D0/DSQRT(SN)
      K = DMIN1(1.0D0D0,SN)
   80 COUNT = 0
      DO 90 I = 1,N
        B(I) = X(I)+K*T(I)
        IF (DABS(B(I)-X(I)).LT.EPS) COUNT = COUNT+1
   90   CONTINUE
C
C     CHECK IF CONVERGED
C
      IF (COUNT.EQ.N) GO TO 150
      CALL LLIKELY(N,B,P)
      IFN = IFN+1
      IF (IFN.GE.NEVALS) GO TO 170
      IF (P.LT.P0-D1*K*TOL) GO TO 100
      K = W*K
      GO TO 80
C
C     NEW LOWEST VALUE
```

```
C
  100 P0 = P
      IG = IG+1
      CALL GLIKELY(N,B,G)
      IFN = IFN+N
C
C     UPDATE HESSIAN
C
      D1 = 0.0D0
      DO 110 I = 1,N
        T(I) = K*T(I)
        C(I) = G(I)-C(I)
  110   D1 = D1+T(I)*C(I)
C
C     CHECK IF +VE DEF ADDITION
C
      IF (D1.LE.0.0D0D0) GO TO 10
      D2 = 0.0D0
      DO 130 I = 1,N
        S = 0.0D0
        DO 120 J = 1,N
  120     S = S+H(I,J)*C(J)
        X(I) = S
  130   D2 = D2+S*C(I)
      D2 = 1+D2/D1
      DO 140 I = 1,N
        DO 140 J = 1,N
  140   H(I,J) = H(I,J)-(T(I)*X(J)+T(J)*X(I)-D2*T(I)*T(J))/D1
      GO TO 40
  150 IFAIL = 0
C SUCCESSFUL CONCLUSION
      RETURN
  160 IFAIL = 1
C N OUT OF RANGE
      RETURN
  170 IFAIL = 2
C TOO MANY FUNCTION EVALUATIONS
      RETURN
  180 IFAIL = 3
C INITIAL POINT INFEASIBLE
      RETURN
 2005 FORMAT( 2X,3G16.4)
      END

      SUBROUTINE FLIKELY(N,TB,F,DRES)
C*****************************************************************
C  This subroutine computes the deviance residuals.
C  Data input: N = dimension of vector BT1;
C            TB = vector combining beta and alpha vectors;
C     Output: DRES = deviance residuals;
C            F = the observed log likelihood function value at BT.
C*****************************************************************
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      INTEGER NOBS,NSTAT,NX,NX1,NVAR,NCOM
      DIMENSION OBS(1000),TIME(1000),XM(1000,8),XM1(1000,8),
     C          Z(1000,5)
      COMMON OBS,TIME,XM,XM1,Z,NOBS,NSTAT,NX,NX1,NVAR,NCOM
      INTEGER N1,N2
      DIMENSION TB(N),B(25),B1(25),BX(5),BX1(5),TEMP(5),DRES(1000)
      N1=(NSTAT-1)*NX
      N2=NSTAT*NVAR+NCOM
      DO 1 I=1,N1
    1 B(I)=TB(I)
      DO 2 I=1,N2
    2 B1(I)=TB(N1+I)
      F=0.0D0
      ONE=1.0D0
      DO 100 I=1,NOBS
C*****************************
C  Loop 20 computes BX(J).  *
C*****************************
      DO 20 J=1,NSTAT-1
      BX(J)=0.0D0
      DO 10 M=1,NX
   10 BX(J)=BX(J)+XM(I,M)*B(M+(J-1)*NX)
   20 CONTINUE
      BX(NSTAT)=0.0D0
      DO 22 J=1,NSTAT
      BX1(J)=0.0D0
C*****************************************************
C  Loop 21 computes BX1(J) for variable coefficient part.  *
```

```
C*****************************************************
      DO 21 M=1,NVAR
      BX1(J)=BX1(J)+XM1(I,M)*B1(M+(J-1)*NVAR)
 21   CONTINUE
 22   CONTINUE
      IF (NCOM.EQ.0) GO TO 25
      DO 24 J=1,NSTAT
      DO 23 M=1,NCOM
      BX1(J)=BX1(J)+XM1(I,M+NVAR)*B1(M+NSTAT*NVAR)
 23   CONTINUE
 24   CONTINUE
 25   CONTINUE
C*****************************************************
      PMAX=BX(1)
      DO 26 J=2,NSTAT
      IF (BX(J).GT.BX(J-1)) PMAX=BX(J)
 26   CONTINUE
      PSUM=0.0D0
      DO 28 J=1,NSTAT
      bx(j)=bx(j)-pmax
      CALL AEXP(BX(J),TEMPP )
      PSUM=PSUM+TEMPP
 28   CONTINUE
      DRES(I)=-DLOG(PSUM)
C*****************************************************
      DO 40 J=1,NSTAT
      IF (BX1(J).LT.0.0D0) GO TO 35
      CALL AEXP(-BX1(J),TEMPP)
      TEMP(J)=BX(J)+(OBS(I)-TIME(I))*BX1(J)-TIME(I)*DLOG(ONE+TEMPP)
      GO TO 40
 35   CALL AEXP(BX1(J),TEMPP)
      TEMP(J)=BX(J)+OBS(I)*BX1(J)-TIME(I)*DLOG(ONE+TEMPP)
 40   CONTINUE
      TEMPMAX=TEMP(1)
      DO 45 J=2,NSTAT
      IF (TEMP(J).GT.TEMP(J-1)) TEMPMAX=TEMP(J)
 45   CONTINUE
      TEMPSUM=0.0D0
      DO 48 J=1,NSTAT
      TEMPP=TEMP(J)-TEMPMAX
      CALL AEXP(TEMPP,TEMP2)
      TEMPSUM=TEMPSUM+TEMP2
 48   CONTINUE
      DRES(I)=DRES(I)+TEMPMAX+DLOG(TEMPSUM)
      F=F+DRES(I)
 100  CONTINUE
      f=-f
      RETURN
      END
```