Statistica Sinica 1(1991), 33-50

TOPICS IN LIKELIHOOD-BASED METHODS FOR LONGITUDINAL DATA ANALYSIS

Nan M. Laird

Harvard University

Abstract: Many popular methods for the analysis of serial measurements obtained in longitudinal studies are based on an underlying multivariate normal distribution with linear mean model for the observations. By modeling the covariance matrix separately from the mean, a broad class of correlation structures can be accommodated. Since the multivariate normal is parameterized only by the mean and covariance of the observations, likelihood-based and moment-based estimation approaches yield similar estimating equations. When the longitudinal responses obtained are categorical, the data structures are similar, but developing flexible model-based approaches which parallel the general linear multivariate normal model is more complex because of two general features of categorical data: the dependence of variance on the mean and the attractiveness of nonlinear models for the mean response. This paper discusses two approaches to modeling these data structures, a general multivariate and a random effects model. We draw parallels with the serial measurements case, and consider the interpretation of the parameters in the model. We discuss maximum likelihood estimation of model parameters under the full likelihood and, for the random effects model, using a conditional likelihood.

Key words and phrases: Repeated categorical response, random effects models, multivariate models for discrete responses.

1. Introduction

Longitudinal studies are commonly undertaken in the health and social sciences for a variety of purposes. Not only do longitudinal studies enable us to eliminate some biases due to selection effects in cross-sectional studies, they also provide greater efficiency for estimating change. A major objective of many studies is to describe growth, aging or time profiles in a response of interest and characterize the causal effects of subject covariates or experimental variables on change over time. The focus of this paper is on probability models and likelihood-based analyses useful in longitudinal studies when the objective is the characterization of these time trends and their dependence on covariates or experimental variables. Other objectives of longitudinal studies may include describing the pattern of association in the responses obtained at different time

points or the temporal sequence of related outcomes. For our purposes, the structure of these associations is not of primary interest, and the parameters describing them will be treated mainly as nuisance parameters.

For purposes of discussing models and methods of analysis it is useful to begin with a simple but flexible design for a longitudinal study, which we will call the standard design. In the standard design, a response is obtained for each subject, indexed by $i = 1, \ldots, n$, at the same set of T time points, indexed by $t = 1, \ldots, T$. The time points need not be equally spaced. Each subject also has a $p \times 1$ vector of covariates, $\boldsymbol{a}_i$ (which may include experimentally manipulated variables), which are assumed constant over time. Thus in the standard design all of the explanatory variables are either pure "within subject" (functions of time or its surrogate) or "between subject" ($\boldsymbol{a}_i$) variables. Further, the within subject variables take on the same values for all subjects since all individuals are measured on the same occasions. This feature of standard designs is useful in characterizing models for longitudinal data.

The standard design is more flexible than it might first appear. For example, it can cover the case where the same subject is measured repeatedly under different experimental conditions. Follinsbee et al. (1988) describe a cross-over experiment where each subject is studied under two experimental conditions in two different periods; serial responses are obtained on each subject under each condition. In this case, both period and time are within subject variables and order of treatment assignment is a between subjects variable.

An epidemiological survey with participants measured annually for the same set of years can be described as a standard design if we take "time" to be calendar year and initial age to be a subject specific covariate. In this setting it is useful to have separate variables for the effects of initial age and age at measurement in order to avoid confounding longitudinal information about age (obtained from within subject changes) with cross-sectional information about age (obtained from subject differences at initial measurement).

Multi-wave or multi-panel longitudinal designs consist of several groups of subjects; within each group there is a standard design, but the time design may vary from group to group. These can be regarded as standard designs with predetermined patterns of missing data.

Although the standard design representation provides a useful way of characterizing many types of longitudinal studies, it will often fail to be sufficiently flexible to describe some. Many epidemiological surveys or studies based on registry data or physician records will have data structures not amenable to any simple characterization. Subjects may have any number of responses taken at any arbitrary set of times, covariates may be fixed for some subjects and change unpredictably for others (e.g. smoking or employment status). Setting aside the

questions of drawing inferences from such observational data, it is useful to have models and methods with sufficient flexibility to handle these types of situations. We refer to these data structures as unbalanced.

Even if the intended design of a longitudinal study is of the standard form, it is often transformed into an unbalanced design by subject attrition or failure to obtain the requisite set of measurements for each subject on each occasion. Because this is such a common problem, especially when dealing with long term studies of human populations, good models and methods of analysis are needed to deal with unbalanced designs that arise as a result of missing data, either deliberate or unintended.

The standard design also permits a simple notation which can be used with both discrete and measured responses. The $T \times 1$ vector of responses is denoted by $\boldsymbol{Y}_i$, where $\boldsymbol{Y}_i^t = (Y_{i1}, \ldots, Y_{iT})$ and the superscript t in $\boldsymbol{Y}_i^t$ means transpose, the $p \times 1$ subject specific covariate vector is $\boldsymbol{a}_i$, and the design on time is specified by the $T \times r$ matrix $\boldsymbol{Z}$, where the rows of $\boldsymbol{Z}$ correspond to the T times and each column of $\boldsymbol{Z}$ contains a suitable function of time, e.g., a constant, a linear trend, a quadratic trend, etc. We remark that $\boldsymbol{Z}$ need not be a function of time; see Lange and Laird (1989) for an experiment involving pressure as the within subject variable.

Following Cox (1972), we distinguish between studies of dependence and studies of association. In studies of dependence, the primary interest usually centers on parameters which model $E(\boldsymbol{Y}_i)$ as a function of time and covariates. This is in contrast to studies of association, where parameters modeling $\mathrm{cov}(\boldsymbol{Y}_i)$ or $E(Y_{it}|y_{it'}, t' < t)$ are of primary interest. For likelihood-based analyses, it is of course necessary to specify fully the entire distribution of $\boldsymbol{Y}_i$, but we will focus our discussion on models which parameterize $E(\boldsymbol{Y}_i)$ as a function of $\boldsymbol{Z}$ and $\boldsymbol{a}_i$, with the covariance parameters being treated primarily as nuisance. The next section summarizes some key features of multivariate normal models for measurements. The remainder of the paper deals with binary response.

2. Linear Models for Serial Measurements

Most of the flexibility of the measured data models stems from the use of the multivariate normal distribution and the use of linear models for the mean parameter vector. The usual distributional assumption for serial measurements is that $\boldsymbol{Y}_i$ is multivariate normal with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. There are three features of the multivariate normal with linear mean structures which make it particularly attractive for the development of a flexible class of models for serial measurements. First, in the absence of covariates, only $T + T(T-1)/2$ parameters are needed for a complete model specification; if T is small relative to n it is not necessary to assume parsimonious models. Second is

the fact that the mean vector and covariance matrix are distinct parameters and may be modeled separately.

When the parameters for the covariance matrix Σ_i are primarily considered nuisance parameters, the purposes of modeling Σ_i include potential efficiency due to estimating a smaller number of parameters (T large) provided we model the structure appropriately, and simpler estimating equations (for certain structures). This must be balanced against a potential loss in efficiency, or inappropriate estimation of the variance of the estimated regression parameters, if we misspecify the structure of Σ_i. If T is small, Σ is often assumed arbitrary but constant for all i. More parsimonious representations for Σ include random effects, autoregressive or other time series, general linear structure, and factor analytic. From the point of view of modeling, the random effects structure is quite flexible in that it can accommodate any degree of imbalance in the data with regard to number and timing of individual responses.

A third property of the multivariate normal model which is attractive for handling longitudinal data is the relation between joint and marginal distributions. When some observations are missing on an individual unit, either at random or completely at random in the sense of Rubin (1976), the contribution to the likelihood for that unit is the multivariate normal kernel of the marginal distribution of the responses that are observed (Little and Rubin(1988), Chapter 7). This marginal distribution has exactly the same form as the joint distribution of the T responses based on a subset of the parameters.

Suppose, for example, unit i is observed on S occasions; let S_i denote an $S \times T$ matrix consisting of the S rows of a $T \times T$ identity matrix corresponding to the S occasions where observations are obtained. Then if we have a model for μ_i and Σ_i in the joint distribution of all T responses, the marginal distribution is multivariate normal with mean $S_i\mu_i$ and covariance matrix $S_i\Sigma_i S_i^t$. In particular, if we assume $\mu_i = X_i\alpha$, the marginal mean is simply $S_iX_i\alpha$. This feature means that in dealing with missing observations, we merely need to change our design matrix to get the appropriate representation for the mean when we use likelihood-based analyses. The ease of handling the covariance parameters depends upon the assumed structure for Σ. In general, $S_i\Sigma S_i^t$ consists of a subset of the rows and columns of Σ. For random effects structures the covariance matrix will retain the same general structure with missing observations. If an autoregressive structure is assumed for Σ, the representation in terms of the autoregressive parameters is more complex. Although computations may be more complex with missing and/or unbalanced data, the equivalence of the representations makes it easy to specify the appropriate likelihood when the data are missing at random, or completely at random.

If in addition to assuming multivariate normality for Y_i, we also assume a

linear structure for μ_i, the ML computations are easily managed using a variety of algorithms. If we have a standard design, and use the model for μ_i suggested by Khatri (1966) and Grizzle and Allen (1969), then we have closed form solutions for the maximum likelihood estimators under wide range of assumptions about the structure for Σ_i. Under this representation, we have

$$\mu_i = Z\psi a_i, \tag{1}$$

where ψ is a $r \times p$ matrix of parameters; each row of ψ corresponds to a time trend (constant, linear, quadratic, etc.). The columns of Z give the effects of covariates on the time parameters. Note that (1) can also be written as a more general linear model

$$\mu_i = X_i\alpha, \tag{2}$$

where X_i is a $T \times q$ matrix and α a $q \times 1$ vector, by setting $X = a_i \otimes Z$, where $\otimes$ denotes direct product, and α is formed by concatenating the rows of μ_i.

With the general linear model and unbalanced designs, the maximum likelihood computations are generally iterative, with the complexity depending upon the structure for Σ_i. A new BMDP release, based on work by Schlucter (1988), handles MLE for this general model, allowing any type of structure for Σ_i. Under the linear model, the MLE of α may also be obtained by iterative generalized least squares (GLS), where the weight matrices are updated at each iteration using the current estimate of Σ_i (see Ware (1985)). This has led to the suggestion of using GLS for α, but method-of-moments type estimations for Σ_i, in an effort to simplify the ML computations; in general, such iterative GLS estimators do not converge to the MLE.

Finally, we note the connection between linear growth curve models and the general longitudinal model. In the growth curve setting, each individual is assumed to have a growth model with a random parameter vector β_i, then the β_i's are modeled as a linear function of the a_i's. With linearity, a growth curve model can be written as a special case of (2) with a random effects structure assumed for Σ_i, implying that the marginal mean of y_i is also a linear function of the covariates, with the same parameters in the model for the β_i's. This generates a class of models sometimes referred to as two-stage random effects models. Diggle (1988) and Azzalini (1987) discuss maximum likelihood for growth curve models which also include an autocorrelation structure for the random error term.

3. Discrete Responses

We now turn to the case where each Y_{it} is categorical. For simplicity we will consider only the case of binary outcomes. The joint distribution of the Y_{it}'s, $t =$

$1,\ldots,T$, is multinomial with 2^T possible outcomes. Notice that this distribution does not have a simple representation in terms of the first and second moments of the Y_{it}'s; rather it assumes a transformation of the Y_{it}'s to a 2^T indicator vector; we let $\boldsymbol{W}_i$ denote this indicator vector and denote the probabilities of its underlying multinomial by $\boldsymbol{\pi}_i^t = \{\pi_{j_1 j_2 \ldots j_T}\}$, where $\mathbf{1}^t\boldsymbol{\pi}_i = 1$. The fully parameterized distribution thus has $2^T - 1$ parameters; this is in contrast to the multivariate normal where the fully parameterized distribution has only $T(T+3)/2$ parameters. Since $2^T - 1$ grows much faster than $T(T+3)/2$, it becomes essential to find parsimonious structures for the parameter set as the number of observations on each individual increases. In this section we discuss some of the difficulties encountered in formulating distributional models for discrete outcomes and some alternative approaches that have been developed.

As in the case with measured responses, the starting point for our work is formulating models for the mean response as a function of time and covariates. When each individual is observed on only one occasion so that we are in the univariate setting, the natural parameterization corresponding to the canonical link function is to assume logit of $E(Y)$ is linear in time and covariates. In the univariate setting the logit model has many attractive features, including shifting the parameter space on (0,1) to the whole real line (see, for example, Cox (1970), Chapter 2) and so it is generally used in the multivariate setting as well. However the nonlinearity of the link function does have drawbacks in the multivariate context, particularly when we seek to develop random effects models. Although other link functions may be used (probit or log-log), any nonlinear one will have many of the same drawbacks as the logit.

The last 25 years have seen the development of logistic and log-linear models for analyzing discrete multivariate outcomes (Birch (1963), Haberman (1974), Bishop et al. (1975)). Although these models are flexible and in widespread use, their utility is restricted to two types of situations: 1) modeling the dependence of a univariate response on a set of predictors or covariates or 2) modeling the association structure between a set of multivariate responses. Because the general log-linear model places structure on the set of joint probabilities of response, $\boldsymbol{\pi}_i$'s, these models are not directly useful for studies of dependence, where interest centers on modeling

$$E(Y_{it}) = P(Y_{it} = 1) = p_{it} = \sum_{j_k \neq t} \pi^i_{j_1 \ldots j_T} \tag{3}$$

as a function of time and covariates. This same point has been made concerning the utility of log-linear models in the context of more general multivariate discrete data (Cox (1972), Prentice (1988), McCullagh and Nelder (1989)).

Following the approach used with measured responses, we write

$$\boldsymbol{\mu}_i = \boldsymbol{Z}\boldsymbol{\psi}\boldsymbol{a}_i \tag{4}$$

for the standard design, or more generally

$$\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\alpha}, \tag{5}$$

where now $\mu_{it} = \text{logit}\, p_{it}$ and $\boldsymbol{\mu}_i^t = (\mu_{i1}, \ldots, \mu_{iT})$. A major obstacle in developing multivariate models for discrete outcomes is that the variance parameters are functions of the mean parameters. In particular,

$$\text{var}(Y_{it}) = p_{it}(1 - p_{it}).$$

In fact, once we assume a parameterization for the marginal mean, we have placed restrictions on the entire set of 2^{T-1} probabilities of the underlying multinomial distribution (equation 3).

The remainder of this section describes two approaches to specifying the joint distribution of the responses, where we take (4) or (5) as the starting point. One parallels the "general multivariate" approach, the other uses a random effects modeling approach.

General multivariate

Recent work by McCullagh and Nelder (1989, Chapter 6.5–6.6), Zhao and Prentice (1989) and Lipsitz et al. (1990) on general multivariate models for discrete multivariate data show promise for application in the longitudinal data setting. Related work on multivariate models for contingency table data appears in Haber (1985). These models use the linear logit model for the vector of marginal probabilities (equation 5), combined with a general structure on the joint probabilities. McCullagh and Nelder's model was developed for the general multivariate case with $T = 2$ where the 2 occasions correspond to different outcomes on the same individual, e.g. presence or absence of two different diseases. Lipsitz et al. also consider the $T = 2$ case, but in the longitudinal data setting. In this case, an individual's probability vector consists of the joint probabilities underlying the 2×2 table obtained by cross-classifying y_{i1} and y_{i2}. The approach is to simply make a 1–1 transformation on the three dimensional parameter space of the joint probabilities to the parameter space consisting of the logits of the two marginal probabilities and a third parameter to make the transformation 1–1. This third parameter can also be thought of as an association (or correlation parameter). McCullagh and Nelder use the log of the odds-ratio of the 2×2

table as the third parameter; Lipsitz et al. also consider the correlation and the risk ratio. Although the correlation is more natural in the measurements setting the odds-ratio allows us to model association without imposing constraints on the margins and has the advantage that asymptotically $\hat{\boldsymbol{\alpha}}$ and the estimated association parameters are independent (McCullagh and Nelder (1989), Palmgren (1989)).

Following Cox (1972), we may write the general log-linear model for the 2×2 case as

$$\ln p(\boldsymbol{y}_i) = \theta_1^i y_{i1} + \theta_2^i y_{i2} + \theta_{12}^i y_{i1} y_{i2} - \Delta^i, \tag{6}$$

where $p(\boldsymbol{y}_i)$ is the joint probability that the random outcome $\boldsymbol{Y}_i = \boldsymbol{y}_i$, and Δ^i is a normalizing constant insuring that the four probabilities $\{\pi_{j_1 j_2}\}$ sum to one. The θ^i parameters have an interpretation as regression parameters in the conditional distribution of y_{i1} given y_{i2}, i.e.

$$\text{logit}\, P(Y_{i1} = 1 | y_{i2}) = \theta_1^i + \theta_{12}^i y_{i2}.$$

Note also that in the $T = 2$ case, $\theta_{12}^i = \ln(\pi_{11}^i \pi_{22}^i / \pi_{12}^i \pi_{21}^i)$. The usual loglinear model would then parameterize the θ^i's as function of covariates (e.g. $\boldsymbol{\theta}_i = \boldsymbol{X}_i \boldsymbol{\alpha}$) for suitable $\boldsymbol{X}_i$, $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}_i^t = (\theta_1^i, \theta_2^i, \theta_{12}^i)$. Instead we transform the parameter space from $\boldsymbol{\theta}_i$ to $\boldsymbol{\eta}_i$, where

$$\eta_{it} = \text{logit}\, p_{it} = \mu_{it}, \quad t = 1, 2$$

and

$$\eta_{i3} = \theta_{12}^i,$$

and make $\boldsymbol{\eta}_i^t = (\eta_{i1}, \eta_{i2}, \eta_{i3})$ a linear function of covariates. Since the transformation is 1–1 we have allowed for arbitrary association structure, but modeled the marginal parameters as functions of the covariates.

This approach has been generalized for the repeated measures setting with arbitrary numbers of observations on each individual by Zhao and Prentice (1989), using the covariances of (y_{it}, y_{ik}) rather than the odds ratios as association parameters. We consider here its application to the longitudinal data setting with T observations on each individual. The general log-linear model for a 2^T table can be expressed as

$$\ln p(\boldsymbol{y}_i) = \sum_{t=1}^{T} \theta_t^i y_{it} + \sum_{k<t} \theta_{kt}^i y_{ik} y_{it} + \sum_{k<m<t} \theta_{kmt}^i y_{ik} y_{im} y_{it}$$
$$+ \ldots + \theta_{123\ldots T}^i \prod_{t=1}^{T} y_{it} - \Delta_i, \tag{7}$$

where $p(\boldsymbol{y}_i) = p(y_{i1}, y_{i2}, \ldots, y_{it})$. The model given in (7) is saturated in parameters; a simple parsimonious model which retains all first order associations is the pairwise model which sets all the θ^i's to zero except for θ^i_t, $t = 1, \ldots, T$ and θ^i_{kt}, $k < t, t = 1, \ldots, T$. In the pairwise model, the θ^i_{kt}'s are again log odds-ratios of the form

$$\theta^i_{12} = \ln(\pi^i_{11j_3\ldots j_T}\pi^i_{22j_3\ldots j_T}/\pi^i_{12j_3\ldots j_T}\pi^i_{21j_3\ldots j_T}), \tag{8}$$

etc. The θ^i_{kt}'s can also be thought of conditional odds-ratios, as they model the association between y_{ik} and y_{it} given all other y_{il}, $l \neq k, t$.

Zhao and Prentice (1989) suggest fixing the third-order and higher θ^i's (θ^i_{kmt}, θ^i_{lkmt} etc.) at predetermined values (setting them equal to zero gives a convenient class of estimators), then making a 1-1 transformation from the lower order θ^i's to the moment parameters $(\boldsymbol{p}_i, \boldsymbol{\Gamma}_i)$, where $\boldsymbol{p}_i$ is the $T \times 1$ vector of marginal moments and $\boldsymbol{\Gamma}_i$ is the $\binom{T}{2}$ vector of marginal correlations,

$$E(y_{ik} - p_{ik})(y_{it} - p_{it})[p_{ik}(1 - p_{ik})p_{it}(1 - p_{it})]^{-1/2}.$$

Then $\boldsymbol{p}_i$ and $\boldsymbol{\Gamma}_i$ are modelled as appropriate functions of the covariates, and parameter vectors $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Solving the likelihood equations gives pseudo-maximum likelihood estimates, and maximum likelihood in the event that the higher-order θ^i's are correctly specified. They also suggest using a robust variance estimator rather than the Fisher Information, which will only be appropriate if the assumed model is correctly specified. We note that the adequacy of the pairwise model can easily be tested using standard log-linear modelling techniques.

Two alternatives to modeling the correlation structure which seem more natural here are to model the log odds-ratios, either the marginal or conditional, as functions of the covariates. The marginal odds ratios, say ω^i_{tk}, are defined as

$$\omega^i_{12} = \ln(\pi_{11+\ldots+}\pi_{22+\ldots+}/\pi_{12+\ldots+}\pi_{21+\ldots+}), \quad \text{etc.},$$

while the conditional odds ratios are the θ^i_{kt}'s, as defined in (8).

Generalizing the case for $T = 2$ and assuming the pairwise model holds, we may write

$$\boldsymbol{\eta}_i = \begin{pmatrix} \boldsymbol{\mu}_i \\ \boldsymbol{\omega}_i \end{pmatrix}$$

using the marginal odds ratios, or

$$\boldsymbol{\eta}^*_i = \begin{pmatrix} \boldsymbol{\mu}_i \\ \boldsymbol{\theta}_i \end{pmatrix}$$

using the conditional (where $\boldsymbol{\theta}_i$ is taken to be the $\binom{T}{2}$ vector of θ^i_{kt}'s) and then set either equal to

$$\begin{pmatrix} \boldsymbol{X}_i\boldsymbol{\alpha} \\ \boldsymbol{Z}_i\boldsymbol{\beta} \end{pmatrix}$$

for appropriate covariate matrices X_i and Z_i. The model using η_i can be considered a special case of McCullagh and Nelder's multivariate model. In the setting where the association parameters are primarily nuisance parameters, we can alternatively use η_i^*. Although using η_i^* may not be as appealing from the point of view of interpretation of the β parameters, it does have the advantage that the θ_i's are invariant to changes in the marginal probabilities, and that the asymptotic covariance of $\hat{\alpha}$ with $\hat{\beta}$ equals $\mathbf{0}$.

This general approach can be extended to incorporate non-zero higher-order parameters, although the number of parameter proliferates rapidly, so that the pseudo-maximum likelihood approach suggested by Zhao and Prentice may be more attractive.

Computation of $(\hat{\alpha},\hat{\beta})$ is fairly straightforward for $T = 2$, using Newton-Rhapson or Fisher Scoring (Lipsitz et al., McCullagh and Nelder). For larger T, the computations become more complicated because evaluation of the likelihood equations requires evaluation of the multinomial probability vector π^i for each i. For the pairwise model (or even for the general model) there is no closed form expression representing π^i as a function of η_i or η_i^* (Bishop, Fienberg and Holland (1975), Ch. 3.4.2). This means one must use a series of iterations for each i (or for each group of individuals with identical covariates), within each step of the Fisher-Scoring or Newton-Rhapson algorithm to compute the π^i. For large n and/or T, the computations thus become prohibitive, although the use of iterative proportional fitting to compute the π^i's may prove to make computations more manageable.

One drawback of the general multivariate model is that the distribution is not reproducible, i.e. for some $s < T$, the marginal distribution of $(y_{i1}, y_{i2}, \ldots, y_{is})$ does not have the same form as (7) with second and higher order associations set to zero, i.e. it is not true that

$$\ln p(y_{i1}, y_{i2}, \ldots, y_{is}) = \sum_{t=1}^{s} \theta_t^i y_{it} + \sum_{k<t=1}^{s} \theta_{kt}^i y_{ik} y_{it} - \Delta^i$$

for general θ^i's. Not only does pairwise independence not hold for the marginal distribution except in special circumstances, but even if it does, the parameters of the corresponding pairwise model are not the same as the θ^i's in the joint model. This means that the appropriate likelihood for the case where we have missing data will be more complicated to compute. Even though model (7) and its pairwise version is a regular exponential family density with sufficient statistics, after reparameterization to (α,β) the likelihood can no longer be expressed in the regular exponential family form, so that the EM algorithm is not a panacea for computing in the missing data case.

We remark that Bonney (1987) has proposed a class of logistic models for longitudinal data, which take the form

$$\text{logit}\,\tilde{p}_{i1} = \boldsymbol{X}_{i1}\boldsymbol{\beta} + \gamma_0$$

and

$$\text{logit}\,\tilde{p}_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \gamma_{t-1}y_{it-1} + \ldots + \gamma_1 y_{i1}, \qquad t > 1,$$

where $\boldsymbol{X}_{it}$ is the tth row of $\boldsymbol{X}_i$, and $\tilde{p}_{it} = P(Y_{it} = 1|y_{it-1}, \ldots, y_{i1})$. Bonney's model is a special case of the general log-linear model, and thus more useful in studies of association rather than dependence, analogous to the time series models for measured data proposed by Rosner, Muñoz et al.(1985). Similar considerations apply to the discrete data models discussed in Rosner (1984) and Connolly and Liang (1988). Neuhaus and Jewell (1989) discuss the relationship between the parameters in these conditional models, mixed models (discussed in the next section) and marginal models.

Two-stage random effects models

Drawing on analogies with linear models for measured data, various authors (Korn and Whittemore (1979); Stiratelli, Laird and Ware (1984); Zeger, Liang and Albert (1988)) have suggested the use of random effects models. This allows the estimation of individual growth curves and rates of change, and also induces a correlation structure in the longitudinal discrete data setting. One variant of this approach patterned after the general random effects model assumes that the Y_{it} are conditionally independent given a vector of "individual" parameters, say $\boldsymbol{b}_i$, with $p^*_{it} = E(Y_{it}|\boldsymbol{b}_i, \boldsymbol{X}_i)$, $\mu^*_{it} = \text{logit}\,p^*_{it}$ and

$$\boldsymbol{\mu}^*_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i. \tag{9}$$

The $\boldsymbol{b}_i$'s are usually assumed to be independently distributed as $N(\boldsymbol{0},\boldsymbol{D})$. Here $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are suitably chosen functions of time and covariates. We use the notation $\boldsymbol{\mu}^*_i$ rather than $\boldsymbol{\mu}_i$ to emphasize the fact that they are logits of different probabilities, p^*_{it} rather than p_{it}. For the standard design (9) may be written

$$\boldsymbol{\mu}^*_i = \boldsymbol{Z}\boldsymbol{\psi}\boldsymbol{a}_i + \boldsymbol{Z}\boldsymbol{b}_i. \tag{10}$$

Zeger, Liang and Albert (1988) use a probit rather than a logit link function and introduce an additional scale parameter for the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{b}_i$. Related mixed models for binary data developed in other contexts are discussed in Wong and Mason (1985), Anderson and Aitkin (1985), Gilmour, Anderson and Rae (1985), McCulloch (1989) and Conaway (1989a).

An important distinction between mixed models for binary data and mixed linear models for measured data is that the conditional and marginal regression parameters are not, in general, equal, i.e. if model (9) holds with $E(\boldsymbol{b}_i) = \mathbf{0}$ so that

$$E(\boldsymbol{\mu}_i^*) = \boldsymbol{X}_i\boldsymbol{\alpha},$$

in general

$$\boldsymbol{\mu}_i = \text{logit}\,[E(\boldsymbol{y}_i|\boldsymbol{X}_i)] \neq \boldsymbol{X}_i\boldsymbol{\alpha}.$$

Zeger et al. show that for the probit link (and the logit via approximation) effects of covariates are greater in magnitude in the random effects model than they are in the marginal regression model. The proof of this phenomenon in general is the same as that which shows there is attenuation in regression parameters for non-linear link models with omitted covariates (Gail et al. (1984)).

There is an extensive literature on parameter estimation with these mixed models, much of it pointing out the computational difficulty in obtaining exact maximum likelihood solutions and suggesting approximations. For the model in (9), the likelihood of the data (sometimes called the marginal likelihood, since we average over the distribution of the $\boldsymbol{b}_i$'s) is

$$L_M = \prod_{i=1}^{n} \int \prod_{t=1}^{T} (p_{it}^*)^{y_{it}} (1 - p_{it}^*)^{1-y_{it}} dF(\boldsymbol{b}_i), \tag{11}$$

where $F(\boldsymbol{b}_i)$ is the multivariate normal distribution of the $\boldsymbol{b}_i$'s. Much of the computational complexity with this model arises because there is no closed form solution for L_M. This led Zeger et al. (1988) and McCulloch (1989) to propose using a probit rather than a logit transformation for p_{it}^*, and Conaway (1989a) to propose a log-log transformation, with a log-gamma distribution for the random effects.

If b_i is a scalar (only one random effect), then the maximum likelihood computations for the mixed model are manageable. The software package EGRET fits the mixed model using a quasi-Newton algorithm, and gives standard error estimates using the inverse sample information matrix. Anderson and Aitken (1985) discuss computation using a Gaussian quadrature approximation to the likelihood. This allows the MLE's to be obtained using iterative reweighted logistic regression. Although this approach is simple to implement, it is limited to relatively small data sets, as the number of "observations" in the logistic regressions is KnT, where K is the number of quadrature points.

An alternative to the full maximum likelihood approach is also available for the case where b_i is a scalar and $\boldsymbol{Z}_i$ is a vector of ones. In this setting with

the standard design, the mixed model is a special case of the Rasch model. The Rasch model of the item analysis literature assumes

$$\mu_{it}^{*} = \phi_t + b_i$$

and arbitrary distributional form for b_i (Rasch (1960)). Here the individual scores, $s_i = \sum_t y_{it} = y_{i+}$, are sufficient statistics for b_i for fixed ϕ_t, thus we may write with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_T)^t$ and $\boldsymbol{b} = (b_1, \ldots, b_n)^t$

$$f(\boldsymbol{y}_i|\boldsymbol{\phi}, b_i) = f(\boldsymbol{y}_i|\boldsymbol{\phi}, s_i) f(s_i|\boldsymbol{\phi}, b_i).$$

This implies the likelihood (for both $\boldsymbol{\phi}$ and $\boldsymbol{b}$) based on an independent sample of individuals with the same $\boldsymbol{\phi}$ factors as

$$L(\boldsymbol{\phi}, \boldsymbol{b}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = L_C(\boldsymbol{\phi}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{s}) L_s(\boldsymbol{\phi}; \boldsymbol{b}, \boldsymbol{s}),$$

where $\boldsymbol{s}^t = (s_1, \ldots, s_n)$, L_C is the likelihood for $\boldsymbol{\phi}$ based on the conditional distribution of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ given $\boldsymbol{s}$ and L_s is the likelihood for $\boldsymbol{\phi}$ and $\boldsymbol{b}$ based on $\boldsymbol{s}$. Thus for an arbitrary distribution, F, on $\boldsymbol{b}$, the marginal likelihood (11) can be written as

$$L_M(\boldsymbol{\phi}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = L_C(\boldsymbol{\phi}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{s}) L_s^{*},$$

where L_s^{*} is the likelihood for $\boldsymbol{\phi}$ based on $\boldsymbol{s}$ only:

$$L_s^{*} \equiv \prod_{i=1}^{n} \int f(s_i|\boldsymbol{\phi}, b_i) dF(b_i).$$

Andersen (1970) has argued that L_s^{*} has very little information about $\boldsymbol{\phi}$ in the absence of any information about F, thus inferences about the item parameters $\boldsymbol{\phi}$ should be based on the conditional likelihood L_C. This is attractive since, as shown by Tjur (1982), the conditional likelihood is Poisson and standard regression packages can be used to estimate $\boldsymbol{\phi}$ and its asymptotic standard error. See Conaway (1989b) for an overview of methods for conditional analysis in this setting.

Before discussing the application of these results to the longitudinal data setting it is worth noting that although the conditional estimates of $\boldsymbol{\phi}$ are consistent and asymptotically nearly efficient, they are not in general equal to the estimates of $\boldsymbol{\phi}$ obtained by maximizing L_M when we assume a parametric form for F. In fact, as with the comparison of marginal and conditional models, there is some attenuation in the estimates of contrasts $(\phi_t - \phi_k)$ using the marginal likelihood.

Intuitively this follows from the fact that the conditional estimator discards data on individuals whose responses at t and k are the same $\{(0,0)$ or $(1,1)\}$ and estimates the $\phi_t - \phi_k$ contrast only from the "changers". Thus the measure of overall (or marginal) change will clearly be less, since it includes data from people who do not change.

We now consider the application of the conditional estimation approach to the longitudinal data setting, starting with a simple case for the standard design with

$$\mu_i^* = Z\psi a_i + Z_1 b_i, \tag{12}$$

where $Z_1^t = (1,1,\ldots,1)$ is the first column of Z and b_i is $N(0,d)$. This is like the compound-symmetry model for measured data, in that it assumes each individual has a unique random "intercept" but the remaining time parameters are fixed. It follows immediately from (12) that $f(y_i|\psi, a_i, b_i)$ can be expressed as

$$f(y_i) = \exp\left\{y_i^t Z\psi a_i + s_{1i} b_i + \sum_{t=1}^{T} \ln(1 - p_{it}^*)\right\}$$

with $s_{1i} = y_i^t Z_1 = y_{i+}$. Since this depends upon y_i only via

$$s_i = y_i^t Z,$$

we write

$$f(y_i) = \exp\left\{s_i^t \psi a_i + s_{1i} b_i + \sum_{t=1}^{T} \ln(1 - p_{it}^*)\right\}.$$

Following Tjur (1982) it is easily seen that

$$f(y_i|s_{1i} = c) = \frac{\exp(s_i^{*t}\psi^* a_i)}{\sum_c \exp(s_i^{*t}\psi^* a_i)},$$

where s_i^* is the $1 \times (r-1)$ vector consisting of $s_{i2},\ldots,s_{ir}$, ψ^* is the $(r-1) \times p$ matrix formed by omitting the first row of ψ, and the $\sum_c$ is over all values of y_i such that $y_{i+} = c$. Thus the likelihood based on the conditional distribution of $y_1,\ldots,y_n$ given $s_1^t = (s_{11},\ldots,s_{1n}) = c^t = (c_1,\ldots,c_n)$ is given by

$$L_C = \frac{\exp(\sum_i^n s_i^{*t}\psi^* a_i)}{\prod_{i=1}^n \sum_{c_i} \exp(s_i^{*t}\psi^* a_i)}. \tag{13}$$

Two things follow immediately from (13): 1) the conditional likelihood does not depend upon the random effects, but there is correspondingly no information about the first row of ψ (i.e. the "main effects") in the conditional likelihood

and 2) conditional estimates of the "time parameters" (ψ^*) can be obtained via an ordinary Poisson regression algorithm.

The attractiveness of the computational simplicity and near optimality of the conditional likelihood approach is offset somewhat by the fact that the conditional likelihood only contains information about the "change" parameters (ψ^*) and not the main effects of time (the first row of ψ). Of course in many (if not all) longitudinal settings the ψ^* are of primary interest, so that this will not be a major limitation. However the model (12) specifies a fairly limited dependence structure which may be inadequate, especially with large T. If we add additional random effects to allow for a more complex dependence structure, we then must condition on additional components of s_i, and the corresponding rows of ψ will drop out of the conditional likelihood. Thus except in fairly simple settings where T is small so that a limited specification on the random effects is adequate, the conditional approach to estimation seems of limited utility and we must resort to using marginal maximum likelihood.

4. Discussion

There exists a growing body of literature on multivariate models for discrete data that have potential application to the longitudinal data setting. A limitation to their usage is the current state of the art in computing, as maximum likelihood computations can be quite formidable. Judging from recent progress, maximum likelihood methods for discrete longitudinal data will eventually be as accessible as those currently available for the measured data setting.

We have limited our discussion to maximum likelihood estimators in this paper. An alternative is to use quasi-likelihood or moment-type estimators, as discussed in Liang and Zeger (1986), Zeger and Liang (1986) and Prentice (1988). Likelihood estimators will be optimal if the assumed model is correct, although there appears to be relatively little loss of efficiency when quasi-likelihood approaches are used. When the model for the mean is correctly specified, but that for the association parameters is not, both the maximum likelihood and the quasi-likelihood estimators are consistent with complete data, although using the standard likelihood approach will misspecify the variance. In this setting, the quasi-likelihood approach, coupled with a robust variance estimator, may be preferable; the relative efficiency of the two in this setting has not been explored.

The situation is somewhat more complex with missing data. Here the quasi-likelihood approach is no longer consistent in general, unless the data are missing completely at random (Rubin (1976)), which is a strong assumption on the missingness process. Lipsitz, Laird and Harrington (1989) derive the bias of the quasi-likelihood estimators with missing at random data for the dichotomous case with $T = 2$. Maximum likelihood is consistent under the correct model speci-

fication, when the data are missing at random, but its properties with model misspecification and missing data have not been studied. Thus in the case of both missing data and model misspecification there is no obvious best choice, especially when we consider that one is often not certain about the model for the missingness. Further research on the relative sensitivity of different estimators to different types of model misspecification and tradeoffs between robustness and efficiency will be necessary to resolve this issue.

Acknowledgement

This work was funded by GM 29745 from the National Institutes of Health.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *J. Roy. Statist. Soc. Ser. B* **32**, 283-301.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *J. Roy. Statist. Soc. Ser. B* **47**, 203-210.

Azzalini, A. (1987). Growth curves analysis for patterned covariance matrices. *New Perspectives in Theoretical and Applied Statistics* (Edited by M. L. Puri, J. P. Vilaplana and W. Wertz), 61-74, John Wiley, New York.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. Ser. B* **25**, 220-233.

Bishop, Y. M. M., Fienberg, S. E. and Holland, R. W. (1975). *Discrete Multivariate Analysis (Theory and Practice)*. The MIT Press, Cambridge.

Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics* **43**, 951-973.

Conaway, M. R. (1989a). Analysis of repeated categorical measurements with conditional likelihood methods. *J. Amer. Statist. Assoc.* **84**, 53-62.

Conaway, M. R. (1989b). A random effects model for binary data. Technical Report, Department of Statistics, The University of Iowa.

Connolly, M. A. and Liang, K.-Y. (1988). Conditional logistic regression models for correlated binary data. *Biometrika* **75**, 501-506.

Cox, D. R. (1970). *Analysis of Binary Data*. Methuen & Co., London.

Cox, D. R. (1972). The analysis of multivariate binary data. *Appl. Statist.* **21**, 113-120.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959-971.

Follinsbee, L. J., McDonnell, W. F. and Horstman, D. H. (1988). Pulmonary function and symptom responses after 6.6-hour exposure to 0.12 ppm ozone with moderate exercise. *J. Air Pollution Control Assoc.* **38**, 28-35.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71**, 431-444.

Gilmour, A. R., Anderson, R. D. and Rae, A. L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**, 593-599.

Grizzle, J. E. and Allen, D. M. (1969). Analysis of growth and dose response curves. *Biometrics* **25**, 357-382.

Haber, M. (1985). Log-linear models for correlated marginal totals of a contingency table. *Comm. Statist. Theory Methods* **14**, 2845-2856.

Haberman, S. J. (1974). *The Analysis of Frequency Data.* University of Chicago Press, Chicago.

Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curve. *Ann. Inst. Statist. Math.* **18**, 75-86.

Korn, E. L. and Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* **35**, 795-802.

Lange, N. and Laird, N. M. (1989). The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *J. Amer. Statist. Assoc.* **84**, 241-247.

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

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1989). Weighted least squares analysis of repeated categorical measurements with outcomes subject to nonresponse. Technical Report, Department of Biostatistics, Harvard University School of Public Health.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1990). Maximum likelihood regression methods for paired binary data. To appear in *Statistics in Medicine.*

Little, R. J. A. and Rubin, D. B. (1986). *Statistical Analysis with Missing Data.* John Wiley, New York.

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

McCulloch, C. E. (1989). Maximum likelihood variance components estimation for binary data. Biometrics Unit Mimeo Series Report, Cornell University.

Neuhaus, J. M. and Jewell, N. P. (1989). Some comments on Rosner's multiple logistic model for clustered data. Technical Report #3, Group in Biostatistics, University of California, Berkeley.

Palmgren, J. (1989). Regression models for bivariate responses. Technical Report #101, Department of Biostatistics, School of Public Health and Community Medicine, University of Washington, Seattle.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Denmarks Paedagogiske Institute, Copenhagen.

Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired data situations. *Biometrics* **40**, 1025-1035.

Rosner, B., Muñoz, A., Tager, I., Speizer, F. and Weiss, S. (1985). The use of an autoregressive model for the analysis of longitudinal data in epidemiologic studies. *Statistics in Medicine* **4**, 457-467.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-590.

Schluchter, M. D. (1988). Analysis of incomplete multivariate data using linear models with structured covariance matrices. *Statistics in Medicine* **7**, 317-324.

Stiratelli, R., Laird, N. and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961-971.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scand. J. Statist.* **9**, 23-30.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *Amer. Statist.* **39**, 95-101.

Wong, G. Y. and Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *J. Amer. Statist. Assoc.* **80**, 513-524.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049-1060.

Zhao, L. P. and Prentice, R. L. (1989). Correlated binary regression using a quadratic exponential model. Technical Report, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center.

Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

(Received November 1989; accepted July 1990)