REGRESSION MODELS FOR CORRELATED BINARY DATA

by

Jennifer S.K. CHAN

Submitted to the University of New South Wales

for the degree of Doctor of Philosophy.

Submitted May, 1996

Contents

Abstract

We develop statistical models to describe the complex events in methadone treatment and study how treatment factors influence the outcome of treatment. The analysis is based on records of drug users who were under methadone maintenance at a single clinic in Western Sydney in 1986. Outcome measures are screens which are recorded as positive or negative for heroin and benzodiazepines use and are measured by urine testing performed once a week. The statistical models used for such problem fit into the context of binary regression. Because more than one drug is tested on each occasion, the problem ultimately becomes a multivariate binary regression problem.

Since the data we consider are repeated measurements over time, they are highly correlated. To account for the serial correlation, we fit two types of fixed effects models: the conditional logistic model via the maximum likelihood approach and the marginal logistic model via the generalized estimating equation approach. Furthermore, to account for the between-subject variation and intra-subject correlation so that we can devise a more patient specific policy for the methadone program, we consider various types of random effects models: the random intercept model using Gaussian quadrature and the method of McGilchrist (1994) for approximating the MLE and the residual MLE; the random coefficients model using the method of Stiratelli, Laird & Ware (1984) and the mixture model using the EM algorithm.

As multiple drug use is common among the methadone clinic patients, we extend our models to bivariate data so that we can study the effect of methadone treatment in reducing multiple drug use. In the model, the logit of the probabilities of both type of drug use as well as their log odds ratios are simultaneously modelled as linear in some covariates and previous outcomes

and this model is further extended to accomodate mixture effects.

We consider primarily the logit link under which e^{β} can be interpreted as an odds ratio. The probit link, however does offer some advantages when it comes to maximum likelihood estimation via the EM algorithm. In the final chapter, we extend the probit-linear mixed model of McCulloch (1994) to allow for correlated random effects.

Acknowledgements

I would like to thank Dr. Anthony Kuk of UNSW and Professor Charles McGilchrist of NCEPH, Australian National University for their guidance and assistance throughout my candidature.

I would also like to thank Dr. James Bell of the Drug and Alcohol Unit, Prince of Wales Hospital for the helpful advice and guidance in the analysis of a large methadone clinic data set which constitutes part of my thesis research.

Finally, I would like to thank my husband and family for their encouragement throughout my candidature.

Declaration

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institute of higher learning, except where due acknowledgement is made in the text.

I also declare that the intellectual content of this thesis is the product of my own work, even though I may have received assistance from others on style, presentation and language expression."

(Signed) _

Chapter 1

Introduction

1.1 Background

In recent years, there has been a resurgence for the support of methadone maintenance programs in many countries as studies have revealed its contribution in reducing the risk of HIV among injecting heroin users in treatment. Associated with this expansion of methadone maintenance, there is a growing research interest in trying to identify the factors that contribute to effective methadone treatment.

A major component of the research is the statistical analysis of a large set of methadone clinic data provided by Dr. James Bell, Director of the Drug and Alcohol Unit, Prince of Wales Hospital. The proposed research is to develop and apply a systematic analytic tool which extracts from the data information in appropriate form in order to evaluate the methadone treatment from various perspectives.

This analysis is based on records of drug users who were under methadone maintenance at a single clinic in Western Sydney in 1986. Outcome measures are drug use of heroin and benzodiazepines as measured by urine testing performed once a week, on a day determined at random. Screens were recorded as positive or negative for each type of drug use. The statistical models used specifically for such problems fit into the context of binary regression since the response variables are the urine tests which result in either a positive or negative outcome. Because more than one drug is tested on each occasion, the problem ultimately becomes a multivariate binary regression problem.

In the context of a generalized linear model, the probability of a positive response, when transformed by a suitable link function ψ, is a linear function of the covariates. In matrix form, the model is

$$\psi(\mathbf{P}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\eta}$ is a $n \times 1$ vector of transformed mean response, $\boldsymbol{\beta}$ a $p \times 1$ vector of fixed effects and $\mathbf{X}$ an $n \times p$ design matrix. Here n represents the total number of observations and p the number of regression coefficients in the model. Some common link functions are the logit link, the probit link and the complementary log-log link. We will mainly use *logit* link under which e^{β} can be interpreted as an odds ratio. The *probit* link, however does offer some advantages when it comes to maximum likelihood estimation via the EM algorithm and so we devote one chapter to the probit link. As the standard logistic distribution is well approximated by a normal distribution with mean 0 and variance $\pi^2/3$, there is not much difference between the two link functions.

It is common that longitudinal responses over time are highly correlated: such problems have been tackled effectively only in the past ten years and methods of analysis are still being developed. Furthermore, the between-

subject variation which induces intra-subject correlation and overdispersion also needs to be accounted for so that we can devise a more patient specific policy for the methadone program.

As multiple drug use is often a common phenomenon among the methadone clinic patients, we extend our models to bivariate data so that we are able to study the effect of methadone treatment in reducing multiple drug use while controlling for their possible interaction.

1.2 The data

The methadone data are restricted to subjects who completed at least four weeks of treatment and those subjects with missing doseage records were excluded. Finally, past experience showed that the treatment was most effective in the first half year of maintenance and beyond that, non-random drop-out began to occur with patients not responding to treatment dropping out which can lead to a false impression of reduced drug use over time. Consequently, our study looked only at results of urine screens collected in the first 26 weeks of treatment. This was done to avoid the distorting effect of patients being on a withdrawal regimen, something that usually began after the first half year of maintenance. The clinic required attendance for dosing seven days per week, with take-home doses of medication only provided in exceptional circumstances.

There were 136 drug users, submitting a total of 2872 urine screens with

16.1% being positive for heroin and 15.6% being positive for benzodiazepines. The doseage averaged over the 2872 incidents is 64mg. For all analyses, each pair of urine screen results rather than each patient served as the unit of analysis.

1.3 Model development

Since our data consists of repeated measurements, it is important to take serial correlation into account in our analysis. In Chapter two, we study and compare models which use two different approaches of modelling: the conditional and unconditional approach. Section 2.1 is mainly devoted to model description. In Section 2.1.1, we study the conditional logistic model proposed by Bonney (1987) in which the outcomes are autoregressed on the previous outcomes. In Section 2.1.2, we study models using the unconditional approach, the marginal logistic model in which the marginal distribution of the outcomes is modelled linearly in some covariates and the association across time of the repeated outcomes for a subject is treated as nuisance and enters only in a working covariance matrix. Parameters are estimated using the generalized estimating equation (GEE) of Liang, Zeger & Qaqish (1992) and Prentice (1988). We make a comparison of these models in Section 2.2. The marginal logistic model is easier to interpret. However, the conditional logistic model has a more tractable likelihood function and can be extended more easily to accommodate random effects. Numerical results are given in Section 2.3.

To account for the between-subject variation and intra-subject correlation,

a number of authors have proposed the use of mixed effects generalized linear model under which the probability of a positive response, when transformed by a suitable link function ψ, is a linear function of fixed as well as random effects. In matrix form,

$$\psi(\mathbf{P}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

where $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_m^T)^T$ is a $qm \times 1$ vector of random effects, $\mathbf{Z}$ the corresponding $n \times qm$ design matrix and m represents the number of subject-specific random vectors of dimension $q \times 1$ each. We usually allow dependence within each $\mathbf{u}_i$ but assume independence between the $\mathbf{u}_i$. In other words, $\mathbf{u}_i$ are iid $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$. In Chapter three, we consider the logit-linear mixed effects models. In addition to accounting for population heterogeneity, over-dispersion and intra-cluster correlation, the incorporation of random effects into the model also enables the pooling of information across different subjects to result in better subject-specific inference as opposed to population-averaged inference.

The presence of random effects, however, complicates the estimation problem considerably. To obtain the marginal likelihood function, one has to integrate out the random effects which except for a few special cases, cannot be performed analytically. For relatively simple problems involving one-dimensional or two-dimensional integrals (Anderson & Aitkin, 1985; Crouch & Spiegelman, 1990), numerical integration can be used to approximate the marginal likelihood. In Section 3.1.1, we fit a random intercept logistic model using Gaussian quadrature as suggested by Anderson & Aitkin (1985) who showed that the maximum likelihood estimators can be obtained by iterative reweighted logistic regression (see equations 3.2 & 3.3) which can be implemented easily using standard packages. We take the number of quadrature points κ to be four. For more accurate determination of the log-likelihood, more quadrature points are needed. Extension to more random components is possible but Gaussian

quadrature becomes infeasible and inaccurate as the dimension of the integral increases.

For more complicated models such as those involving crossed random effects, the marginal likelihoods involve high dimensional integrals which are beyond the scope of numerical integration; see Karim & Zeger (1992) for two models where the integrals involved are of dimensions 20 and 40 respectively. The intractability of the likelihood function has led various authors to propose a host of alternative estimation methods rather than carrying out maximum likelihood estimation exactly. These include the approximate maximum likelihood and approximate residual maximum likelihood estimators proposed by Schall (1991), McGilchrist (1994) and Drum & McCullagh (1993); the penalized quasi-likelihood approach of Breslow & Clayton (1993); the Gibbs sampling Bayesian approach of Zeger & Karim (1991); the estimating function approach of Waclawiw & Liang (1993) and the iterative bias correction approach of Kuk (1995). In Section 3.1.2, we consider the random intercept model using the estimating techniques of Schall (1991) and McGilchrist (1994).

McGilchrist (1994) suggested maximizing the penalised log-likelihood function which can also be interpreted as the complete data log-likelihood function based on the observed outcomes $\mathbf{Y}$ and the random components $\mathbf{u}$. A Bayesian interpretation of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ is that they are the posterior mode under a diffuse prior for $\boldsymbol{\beta}$ and a normal prior for $\mathbf{u}$.

However, for models with correlated random effects, there is no explicit formula for $\widehat{\boldsymbol{\Sigma}}$ when $\boldsymbol{\Sigma}$ is not diagonal. In Section 3.1.3, we use the method of Stiratelli, Laird & Ware (1984) for the random coefficients model with correlated random effects. Following Stiratelli et al, we assume a diffuse prior

for $\boldsymbol{\beta}$ in addition to the $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ prior for $\mathbf{u}_i$. We can then estimate $\boldsymbol{\Sigma}$ by an EM type estimate which requires $\mathrm{E}(\mathbf{u}_i|\mathbf{Y})$ and $\mathrm{Cov}(\mathbf{u}_i|\mathbf{Y})$. Stiratelli et al suggested approximating the posterior distribution of $(\boldsymbol{\beta}, \mathbf{u})$ given $\mathbf{Y}$ by a normal distribution that has the same mode and curvature as the true posterior. In other words, we can replace the posterior mean $\mathrm{E}(\mathbf{u}_i|\mathbf{Y})$ by the posterior mode $\hat{\mathbf{u}}_i$. The conditional covariance matrix $\mathrm{Cov}(\mathbf{u}_i|\mathbf{Y})$ can be approximated by inverting the Hessian matrix of the log posterior density of $(\boldsymbol{\beta}, \mathbf{u})$ given $\mathbf{Y}$ which is equivalent to the Hessian matrix of the complete data log-likelihood because of the diffuse prior assumption for $\boldsymbol{\beta}$.

In Chapter four, we consider the mixture model which can be regarded as a discrete random effects model. The model postulates that there are two or more groups of patients who react differently to methadone treatment. This type of model is more tractable than the random coefficients model as the likelihood function can be computed easily without the need of integration. Estimation can be carried out using the EM algorithm and model selection can be based on the Akaike Information Criterion(AIC).

The model selected is a 3 group mixture model. Under this model, methadone treatment leads to cessation of heroin use for about 30% of the subjects regardless of the methadone doses used. Another 46% of the subjects responded to treatment in a dose-dependent manner with reduced heroin use at high doses of methadone. The remaining 24% of the patients failed to respond to treatment in this study. These findings are consistent with clinical experience.

As multiple drug use is often a common phenomenon among the methadone clinic patients, in Chapter five, we devise a model that enables us to study the effect of methadone treatment in reducing multiple drug use, say heroin

and benzodiazepines while controlling for their possible interaction. In Section 5.1.1, we consider a model in which the logit of the probabilities of both types of drug use as well as their log odds ratio are simultaneously modelled as linear in some covariates. The serial correlation within subject is accounted for by including the previous outcomes of both drugs and their interaction as covariates. Clustered data, as opposed to longitudinal bivariate binary data, has been analysed by Lefkopoulou, Moore & Ryan (1989). Unlike the GEE approach of Zeger & Liang (1991) and the pseudolikelihood approach of Liang & Zeger (1989), the proposed model has a tractable likelihood and so a full likelihood analysis is possible. It can also be easily extended to incorporate mixture or random effects. The Newton-Raphson method is used to obtain the maximum likelihood estimates. For the methadone data, an interesting finding is that the odds ratio seems to depend on the previous outcomes for heroin and benzodiazepines only through whether they are concordant or discordant. This suggests that the strength of the present association between the two drugs depends on the strength of the association last week.

In Section 5.1.2, we consider the bivariate mixture model which postulates that there are two groups of patients who react differently to methadone treatment. Estimation is carried out using the EM algorithm and model selection is based on the Akaike Information Criterion (AIC).

McCulloch (1994) pointed out several advantages of using the probit link instead of the customary logit link. For example, the probit link function is preserved when modelling the marginal distribution. Moreover, by viewing a probit-normal model as a threshold model that results from dichotomizing some unobserved continuous outcomes from a Gaussian mixed model, it becomes feasible to use the EM algorithm to find the maximum likelihood estimates. McCulloch considered only independent random effects. In Chapter

six, we extend McCulloch's model by allowing correlated random effects. This extension widens the applicability of the model considerably.

In Section 6.1, we introduce the probit-linear mixed model with correlated random effects and the applicability of the model is demonstrated by quoting four useful models as special cases of the model. In Section 6.1.1, we give a detailed description of the maximum likelihood estimation of the fixed effects $\boldsymbol{\beta}$ and the variance components $\boldsymbol{\Sigma}_r$ by the EM algorithm. Because of the probit link assumption, simplification to the EM algorithm can be made. The derivation is similar to that of McCulloch (1994) but we end up with a slightly different formula for the M-step. McCulloch's procedure is closely related to version 1 of the ECME algorithm proposed by Liu & Rubin (1994, P.641). In Section 6.1.2, we describe a Monte Carlo implementation of the E-step of the algorithm via Gibbs sampling. In Section 6.1.3, we consider the estimation of standard errors. When the E-step requires Monte Carlo method, the SEM algorithm (Meng & Rubin, 1991) for calculating standard errors is numerically unstable and extremely computing intensive and so we resort to inverting a Monte Carlo estimate of the information matrix. In Section 6.1.4, we describe a method for accounting the Monte Carlo variation explicitly. In Section 6.2, we illustrate the flexibility and feasibility of our methods by fitting two models to the salamander mating data reported in McCullagh & Nelder (1989, pp. 439-450). Both models assume a probit link and include the male species, the female species, their interaction and season (fall versus summer) as fixed effects. In Section 6.2.1, we fit model 1 which includes male and female animal effects as random effects. As the same animals were used in the first two mating experiments, the effects of the same animal over the two occasions are correlated. In Section 6.2.2, we fit model 2 in which the random effects are classified by species as well as by gender. Finally, in Section 6.3, we consider

some useful extensions of the model. In Section 6.3.1, we extend the model to allow correlated errors as well as correlated random effects and a brief description of the model that allows for autocorrelated errors together with the estimation procedure are given. In Section 6.3.2, we extend the model to multivariate clustered binary data which arise for example in the study of multiple binary traits in animal breeding (Foulley, Gianola & Im, 1989) and the estimation procedure is described briefly.

1.4 The covariates

To identify factors associated with drug use, there has been a substantial body of evidence that methadone *dose* is important in influencing continued drug use and so it is necessary to take into account the fluctuations of methadone dose in assessing the influence of other treatment factors. As a result, methadone dose is included in the models as one covariate. Another covariate included is the duration of treatment in weeks called *time*. Since it is unrealistic to expect a linear time trend over a very long period of time, we take logarithm of the time. Note that in fitting the conditional AR(1) model with the logarithm of time as covariate instead of time, we end up with a lower AIC value of 2090.02 as compared to 2097.44 for model with time as covariate. As a result, we take logarithm of the time, log$t$ for the time effect in all our subsequent models. We also include the interaction effect between dose and time, the product term $d_{it} \times \log t$ as covariate initially but it is not significant and is dropped subsequently.

Since longitudinal responses over time are highly correlated, the serial correlation within subject is accounted for in the conditional model by including another covariate, the *previous outcome* into the model. For simplicity of modelling, we consider autoregressive models up to AR(2). For mixture models and bivariate models, we only consider AR(1) models.

Chapter 2

Marginal and conditional logistic models

2.1 The models

Let $Y_{it}(i = 1, \ldots, m;\ t = 1, \ldots, n_i)$ denote the observed outcome of heroin use of the i-th patient at time t and $n = n_1 + \ldots + n_m$ the total number of observations. We shall consider models where the marginal probabilities $P_{it} = \Pr(Y_{it} = 1)$ or the conditional probabilities $\Pr(Y_{it} = 1|Y_{i1}, \ldots, Y_{i,t-1})$ are logit-linear in some covariates.

2.1.1 Marginal logistic model

The main focus of this model is on the marginal distribution of Y_{it}. It is assumed that

$$\text{logit}(P_{it}) = \eta_{it} = \beta_o + \beta_d \, d_{it} + \beta_t \, \text{log}t \tag{2.1}$$

where d_{it} is the doseage administered to patient i at time t. The association across time between the repeated outcomes for a subject is treated as nuisance and is entered only in a working correlation matrix that appears in the estimating equation. An advantage of this approach is that the resulting estimates are robust to misspecification of the correlation structure provided the mean model is correctly specified. The working correlation matrix that we assume for patient i is

$$\Psi_i(\rho) = \begin{pmatrix} 1 & \rho & \dots & \rho^{n_i-1} \\ \rho & 1 & \dots & \rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \dots & 1 \end{pmatrix} \tag{2.2}$$

which corresponds to an autoregressive process of order 1. The generalized estimating equation (Prentice, 1988) of the model is

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{m} \frac{\partial \mathbf{P}_i^T}{\partial \boldsymbol{\beta}} \, \text{Cov}^{-1*}(\mathbf{Y}_i) \, (\mathbf{Y}_i - \mathbf{P}_i) = \mathbf{0} \tag{2.3}$$

where

$$\frac{\partial \mathbf{P}_i^T}{\partial \boldsymbol{\beta}} = \mathbf{X}_i^T \, \text{Diag}(S_{i1}^2, \dots, S_{in_i}^2), \tag{2.4}$$

$\mathbf{X}_i$ is the $n_i \times 3$ matrix of patient i with row vectors $(1, d_{it}, \log t)$, $t = 1, \ldots, n_i$, $\mathbf{P}_i = (P_{i1}, \ldots, P_{in_i})^T$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$. We set the working covariance matrix of $\mathbf{Y}_i$ to be

$$\text{Cov}^*(\mathbf{Y}_i) = \mathbf{V}_i = \text{Diag}(S_{i1}, \ldots, S_{in_i})\ \Psi_i(\rho)\ \text{Diag}(S_{i1}, \ldots, S_{in_i})$$

where $S_{it}^2 = P_{it}(1 - P_{it}) = e^{\eta_{it}}/(1 + e^{\eta_{it}})^2$. We estimate ρ by

$$\widehat{\rho} = \frac{1}{n-m} \sum_{i=1}^{m} \sum_{t=1}^{n_i - 1} \frac{(Y_{i,t+1} - P_{i,t+1})(Y_{it} - P_{it})}{[P_{i,t+1}\ (1 - P_{i,t+1})\ P_{it}\ (1 - P_{it})]^{\frac{1}{2}}}. \tag{2.5}$$

With estimated $\rho^{(k)}$, we can update $\boldsymbol{\beta}$ to $\boldsymbol{\beta}^{(k+1)}$ by solving (2.3) using the Newton-Raphson method. Then ρ is subsequently updated using $\boldsymbol{\beta}^{(k+1)}$ in (2.5) and the cycle repeats again until convergence is reached. Finally, the variance-covariance matrix of $\boldsymbol{\beta}$ (Prentice, 1988) is

$$\begin{aligned} \text{Cov}(\boldsymbol{\beta}) \;=\; & \Big[\sum_{i=1}^{m} \Big(\frac{\partial \mathbf{P}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} \frac{\partial \mathbf{P}_i}{\partial \boldsymbol{\beta}^T}\Big)\Big]^{-1} \Big[\sum_{i=1}^{m} \Big(\frac{\partial \mathbf{P}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} \text{Cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} \frac{\partial \mathbf{P}_i}{\partial \boldsymbol{\beta}^T}\Big)\Big] \\ & \Big[\sum_{i=1}^{m} \Big(\frac{\partial \mathbf{P}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} \frac{\partial \mathbf{P}_i}{\partial \boldsymbol{\beta}^T}\Big)\Big]^{-1} \end{aligned} \tag{2.6}$$

where $\text{Cov}(\mathbf{Y}_i)$ is estimated by $(\mathbf{Y}_i - \widehat{\mathbf{P}}_i)(\mathbf{Y}_i - \widehat{\mathbf{P}}_i)^T$.

2.1.2 Conditional logistic model

In this model, the serial correlation within subject is accounted for by including the previous outcomes as covariates. For AR(1) model, it is assumed that $\Pr(Y_{it} = 1 | Y_{i1}, \ldots, Y_{i,t-1}) = \Pr(Y_{it} = 1 | Y_{i,t-1})$ and

$$\text{logit}[\Pr(Y_{it} = 1|Y_{i,t-1})] = \eta_{it} = \beta_o + \beta_d\, d_{it} + \beta_t\, \log t + \beta_{p1}\, Y_{i,t-1} \tag{2.7}$$

and for AR(2) model, it is assumed that

$$\text{logit}[\Pr(Y_{it} = 1|Y_{i,t-1},Y_{i,t-2})] = \eta_{it} = \tag{2.8}$$
$$\beta_o + \beta_d\, d_{it} + \beta_t\, \log t + \beta_{p1}\, Y_{i,t-1} + \beta_{p2}\, Y_{i,t-2}.$$

The likelihood function is

$$\prod_{i=1}^{m} \Pr(Y_{i1}, \ldots, Y_{in_i}) = \prod_{i=1}^{m} \prod_{t=1}^{n_i} \Pr(Y_{it}|Y_{i,t-1}, Y_{i,t-2}) = \prod_{i=1}^{m} \prod_{t=1}^{n_i} \left(\frac{e^{Y_{it}\eta_{it}}}{1 + e^{\eta_{it}}}\right) \tag{2.9}$$

and the log-likelihood function is

$$l(\boldsymbol{\beta}; \mathbf{Y}) = \sum_{i=1}^{m} \sum_{t=1}^{n_i} Y_{it}\eta_{it} - \sum_{i=1}^{m} \sum_{t=1}^{n_i} \log(1 + e^{\eta_{it}}). \tag{2.10}$$

Bonney (1987) showed that we can obtain the maximum likelihood estimates for $\boldsymbol{\beta}$ using a standard logistic regression procedure with a suitably augmented design matrix. The variance-covariance matrix of $\boldsymbol{\beta}$ can be obtained by inverting $-l''(\boldsymbol{\beta})$.

2.2 Comparsion of marginal and conditional logistic models

The marginal logistic model is a model for the marginal distribution of Y_{it} and the parameters of such a model can be interpreted easily as the marginal log odds ratios. Moreover, the GEE approach requires only the specification of a working correlation matrix which is not necessarily the true correlation matrix. On the other hand, since the association structure is treated as nuisance, the marginal model is not very helpful if the association structure is of interest. Furthermore, a full likelihood approach to inference is not possible under a marginal logistic model as the marginal distributions alone do not determine the joint distribution completely.

The conditional logistic model conditions on the previous responses which complicates the interpretation of the parameters. The regression coefficients $\boldsymbol{\beta}$ can be interpreted as log odds ratios only conditionally but not unconditionally. However, the conditional logistic model accounts for the dependence between Y_{it} and $Y_{i,t-1}$ explicitly and so is preferable if such dependence is of interest. Another advantage of the conditional logistic model is that its likelihood function can be written down quite easily as in (2.9) which facilitates likelihood inference. Also, it is quite easy to extend the conditional logistic model by incorporating random effects to the right hand side of (2.7) or (2.8). Such a random effects model is useful in accounting for population heterogeneity and in subject specific inference.

The marginal and conditional logistic models are incompatible in that if the marginal distributions are logistic, then the conditional distributions are not and vice versa. In a recent development, Azzalini (1994) completed the

specification of a marginal logistic model by modelling the pairwise odds ratio between Y_{it} and $Y_{i,t-1}$ in addition to their marginal distributions. As a result, it is possible to write down the full likelihood function and so maximum likelihood estimators can be obtained. However, the likelihood contribution from each patient as given by equations (5), (6) and (7) of Azzalini's paper is quite complicated and is not as tractable as the likelihood (2.9) under the conditional logistic model. For this reason, we do not use Azzalini's model or its random effects extension to analyse the methadone clinic data.

2.3 Numerical results

The results of fitting the marginal and conditional logistic AR(1) models to the methadone clinic data are given in Table 1. The result for conditional AR(2) model is given in Table 2. We see that the results for AR(1) and AR(2) models are similar and the AR(2) model is preferable in terms of AIC (2090.02 for AR(1) model and 2032.91 for AR(2) model). Since the dose and time effects of both models are significant, our conclusion is that reduced heroin use is associated with increased methadone dose and increased duration in treatment. There is a strong positive association between the present and previous outcomes. In fact, some patients in treatment tend to use heroin heavily while others do not. We have also considered the interaction between dose and time by including the product term $d_{it} \times \log t$ as a covariate. However, the interaction effect is not significant (p-value= 0.59 under marginal model with AR(1) working correlation; p-value= 0.46 under conditional AR(1) model). As a result, we do not include interaction between dose and time in our subsequent

analyses.

The results of fitting a conditional AR(1) model to each patient are given in Table 3. Only 24 patients have convergent estimates and there appears to be substantial variation between these estimates. For formal testing, we extend the score test of homogeneity (Commenges et al, 1994) to the case of the conditional logistic model. More precisely, we are using the score test to test model (2.7) as the null hypothesis against the alternative that one of the β's in (2.7) is random. We conduct the score test for each covariate separately and the results are given in Table 4. It can be seen that all tests are significant with the test for the random intercept being most significant.

Chapter 3

Random effects model

3.1 The models

As there is strong evidence that the regression coefficients are patient-specific, we first incorporate a random intercept to the conditional logistic model (2.7). The extended model is

$$\text{logit}[\Pr(Y_{it} = 1|Y_{i,t-1})] = \eta_{it} = \beta_o + u_i + \beta_d\, d_{it} + \beta_t\, \log t + \beta_{p1}\, Y_{i,t-1} \quad (3.1)$$

where the u_i's are independent and identically distributed as $\mathcal{N}(0, \sigma^2)$. This model can be written in matrix form as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \quad (3.2)$$

where $\boldsymbol{\beta} = (\beta_o, \beta_d, \beta_t, \beta_{p1})^T$, $\mathbf{X}$ the corresponding $n \times 4$ design matrix, $\mathbf{u} = (u_1, \ldots, u_m)^T$ the patient effects and $\mathbf{Z}$ the corresponding $n \times m$ design matrix

for $\mathbf{u}$.

We have seen from the results of fitting the conditional AR(1) model separately to each patient and the score tests that apart from the intercept, some other regression coefficients may also be random. Therefore, we go beyond the random intercept model by including other random coefficients to (3.1). For example, by setting all regression coefficients to be random, the model becomes

$$\text{logit}[\Pr(Y_{it} = 1|Y_{i,t-1})] = \eta_{it} =$$
$$\beta_o + u_{oi} + (\beta_d + u_{di})\, d_{it} + (\beta_t + u_{ti})\, \text{log}t + (\beta_{p1} + u_{p1,i})\, Y_{i,t-1}. \qquad (3.3)$$

We can also express this model in matrix form as in (3.2) where $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)^T$ a $4m \times 1$ vector of the patient effects and $\mathbf{Z}$ the corresponding $n \times 4m$ design matrix for $\mathbf{u}$.

3.1.1 Random intercept model using MLE and Gaussian quadrature

We define the model as in (3.1) and let $u_i = \sigma u_i^*$ where u_i^* has a standard normal distribution. The log-likelihood function is

$$l(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{m} \log\Big\{ \int_{-\infty}^{\infty} \Big(\prod_{t=1}^{n_i} \frac{e^{Y_{it}(\mathbf{x}_{it}\beta + u_i^*\sigma)}}{1 + e^{\mathbf{x}_{it}\beta + u_i^*\sigma}} \Big) \phi(u_i^*) du_i^* \Big\} \qquad (3.4)$$

where $\phi(u_i^*)$ is the standard normal probability density function. Using Gaussian quadrature (Anderson & Aitkin, 1985), a numerical approximation of the log-likelihood function is

$$l(\boldsymbol{\beta}, \sigma) \simeq \sum_{i=1}^{m} \log\Big\{ \sum_{v=1}^{\kappa} \Big(\prod_{t=1}^{n_i} \frac{e^{Y_{it}(\mathbf{x}_{it}\beta + \Lambda_v\sigma)}}{1 + e^{\mathbf{x}_{it}\beta + \Lambda_v\sigma}} \Big) A_v \Big\} \tag{3.5}$$

where Λ_v are the Gaussian quadrature points and A_v, the associated weight factors. The terms $\sqrt{\pi}A_v$ and $\Lambda_v/\sqrt{2}$ are given in Abramowitz and Stegun (1972, p. 924). Differentiating (3.5) with respect to $\beta_j, j = 1, \ldots, 4$ and σ, we obtain

$$\sum_{v=1}^{\kappa}\sum_{i=1}^{m}\sum_{t=1}^{n_i} \frac{x_{itj}\Big[\prod_{t=1}^{n_i}\Big(\frac{e^{Y_{it}(\mathbf{x}_{it}\beta+\Lambda_v\sigma)}}{1+e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}\Big)\Big]A_v}{\sum_{v=1}^{\kappa}\Big[\prod_{t=1}^{n_i}\Big(\frac{e^{Y_{it}(\mathbf{x}_{it}\beta+\Lambda_v\sigma)}}{1+e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}\Big)\Big]A_v}\Big(Y_{it} - \frac{e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}{1+e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}\Big) = 0, \tag{3.6}$$

$j = 1, \ldots, 4$ and

$$\sum_{v=1}^{\kappa}\sum_{i=1}^{m}\sum_{t=1}^{n_i} \frac{\Lambda_v\Big[\prod_{t=1}^{n_i}\Big(\frac{e^{Y_{it}(\mathbf{x}_{it}\beta+\Lambda_v\sigma)}}{1+e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}\Big)\Big]A_v}{\sum_{v=1}^{\kappa}\Big[\prod_{t=1}^{n_i}\Big(\frac{e^{Y_{it}(\mathbf{x}_{it}\beta+\Lambda_v\sigma)}}{1+e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}\Big)\Big]A_v}\Big(Y_{it} - \frac{e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}{1+e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}\Big) = 0. \tag{3.7}$$

Define $Y_{itv} = Y_{it}$, $\mathbf{x}_{itv} = (\mathbf{x}_{it}, \Lambda_v)$ and

$$w_{itv} = w_{iv} = \frac{\Big[\prod_{t=1}^{n_i}\Big(\frac{e^{Y_{it}(\mathbf{x}_{it}\beta+\Lambda_v\sigma)}}{1+e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}\Big)\Big]A_v}{\sum_{v=1}^{\kappa}\Big[\prod_{t=1}^{n_i}\Big(\frac{e^{Y_{it}(\mathbf{x}_{it}\beta+\Lambda_v\sigma)}}{1+e^{\mathbf{x}_{it}\beta+\Lambda_v\sigma}}\Big)\Big]A_v}, \tag{3.8}$$

equations (3.6) and (3.7) are those for a weighted logistic regression where the weights w_{itv} depend on the parameters being estimated. This suggests the use of iteratively reweighted logistic regression analysis. The resulting method can be implemented fairly easily as each weighted logistic regression can be done using any of the standard statistical packages. We use the *SAS* procedure

LOGISTIC to perform the logistic regression and write a *MACRO* to do the iterative reweighting.

A complete description of the iterative procedures is as follow:

1. Choose $\sigma = 0$ and carry out an unweighted logistic regression to obtain an initial estimate of $\boldsymbol{\beta}$.

2. Expand the vector of the dependent variable $\mathbf{Y}$ by defining $Y_{itv} = Y_{it}$. Note that the resulting vector is of length $n \times \kappa$ where n denotes the total number of observations and κ the number of Gaussian quadrature points. Also define $\mathbf{x}_{itv} = (\mathbf{x}_{it}, \Lambda_v)$.

3. Calculate $U_{itv} = \exp[Y_{it}(\mathbf{x}_{it}\boldsymbol{\beta} + \Lambda_v\sigma)] \;/\; [1 + \exp(\mathbf{x}_{it}\boldsymbol{\beta} + \Lambda_v\sigma)]$ using $\sigma = 1$ and $\boldsymbol{\beta}$ from step 1.

4. Calculate $A^*_{iv} = \left(\prod_{t=1}^{n_i} U_{itv}\right) A_v$ and $L_i = \sum_{v=1}^{\kappa} A^*_{iv}$ to create the weights $w_{itv} = A^*_{iv}/L_i$.

5. Use these weights to perform a weighted logistic regression to re-estimate $\boldsymbol{\beta}$ and σ.

6. Recalculate U_{itv}.

Then we iterate step 4, 5 and 6 until convergence is reached. It is obvious that κ should be as small as practicable. Bock and Aitkin (1981) reported that $\kappa = 2$ or $\kappa = 3$ was sufficient. For more accurate determination of the log-likelihood, more quadrature points are needed, say $\kappa = 5$. For models involving more than one random component, say two random components, the extended vector $\mathbf{Y}$ has length $n \times \kappa_1 \times \kappa_2$ which is enormous for the methadone data. As the result, this method of estimation becomes practically infeasible.

3.1.2 Random intercept model using approximate MLE and approximate REML

With the random intercept model as in (3.1), the log probability density function for $\mathbf{u}$ is

$$l_u(\mathbf{u};\sigma^2) = -\frac{1}{2}\sum_{i=1}^{m}\left[\log(2\pi\sigma^2) + \sigma^{-2}u_i^2\right]. \tag{3.9}$$

McGilchrist (1994) suggested maximizing the sum of (2.10) and (3.9) with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ to obtain $\widehat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$, the so-called best linear unbiased predictors or penalised likelihood estimators if l_u is regarded as a penalty function. The sum of (2.10) and (3.9) can also be interpreted as the complete data log-likelihood function based on $\mathbf{Y}$ and $\mathbf{u}$. A Bayesian interpretation of $\widehat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ is that they are the posterior mode under a diffuse prior for $\boldsymbol{\beta}$ and a normal prior for $\mathbf{u}$. The Newton-Raphson step for finding $\widehat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ is

$$\begin{bmatrix} \boldsymbol{\beta}^{(k+1)} \\ \mathbf{u}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}^{(k)} \\ \mathbf{u}^{(k)} \end{bmatrix} + \begin{bmatrix} \mathbf{X}^T\mathbf{S}^2\mathbf{X} & \mathbf{X}^T\mathbf{S}^2\mathbf{Z} \\ \mathbf{Z}^T\mathbf{S}^2\mathbf{X} & \mathbf{Z}^T\mathbf{S}^2\mathbf{Z} + \sigma^{-2}\mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^T(\mathbf{Y}-\mathbf{P}) \\ \mathbf{Z}^T(\mathbf{Y}-\mathbf{P}) - \sigma^{-2}\mathbf{u}^{(k)} \end{bmatrix} \tag{3.10}$$

where $\mathbf{P} = (P_{11},\ldots,P_{1n_1},P_{21},\ldots,P_{mn_m})$, $P_{it} = e^{\eta_{it}}/(1+e^{\eta_{it}})$, $\mathbf{S}^2 = \mathrm{Diag}(S_{11}^2,\ldots,S_{1n_1}^2,S_{21}^2,\ldots,S_{mn_m}^2)$ and $S_{it}^2 = e^{\eta_{it}}/(1+e^{\eta_{it}})^2$. By writing

$$\begin{bmatrix} \mathbf{X}^T\mathbf{S}^2\mathbf{X} & \mathbf{X}^T\mathbf{S}^2\mathbf{Z} \\ \mathbf{Z}^T\mathbf{S}^2\mathbf{X} & \mathbf{Z}^T\mathbf{S}^2\mathbf{Z} + \sigma^{-2}\mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \cdot & \cdot \\ \cdot & \mathbf{T} \end{bmatrix} \tag{3.11}$$

and $\mathbf{T}^* = (\mathbf{Z}^T\mathbf{S}^2\mathbf{Z} + \sigma^{-2}\mathbf{I})^{-1}$, McGilchrist (1994) proposed updating σ^2 by

$$\widehat{\sigma}^2_{ML} = \frac{\mathbf{u}^T\mathbf{u}}{m - \text{tr}(\mathbf{T}^*)/\sigma^2} \tag{3.12}$$

and

$$\widehat{\sigma}^2_{REML} = \frac{\mathbf{u}^T\mathbf{u}}{m - \text{tr}(\mathbf{T})/\sigma^2}, \tag{3.13}$$

where all the quantities on the right hand side of (3.12) and (3.13) are evaluated at the current estimates of $\boldsymbol{\beta}$, $\mathbf{u}$ and σ^2. Using this new estimate of σ^2, we can update our estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$ by (3.10). McGilchrist (1994) suggested iterating (3.10) and (3.12) or (3.13) until convergence to obtain approximate maximum likelihood (ML) or residual maximum likelihood (REML) estimates.

3.1.3 Random coefficients model

Suppose we have q random coefficients in (2.7). Let $\mathbf{b}_i$, $i = 1, \ldots, m$ with mean $\boldsymbol{\beta}_1$ denote the patient-specific regression coefficients and $\boldsymbol{\beta}_2$ the non-patient-specific regression coefficients. It is assumed that $\mathbf{b}_1, \ldots, \mathbf{b}_m$ are i.i.d. $\mathcal{N}_q(\boldsymbol{\beta}_1, \boldsymbol{\Sigma})$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ and $\mathbf{u}_i = \mathbf{b}_i - \boldsymbol{\beta}_1$. We can express the model in matrix form as (3.2) where $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_m^T)^T$ is now a $qm \times 1$ vector of random effects and $\mathbf{Z}$ the corresponding $n \times qm$ design matrix.

The estimation of $\boldsymbol{\beta}$ and $\mathbf{u}$ follows exactly as in (3.10) with $\sigma^{-2}\mathbf{I}$ and $\sigma^{-2}\mathbf{u}$ replaced by $\mathbf{I} \otimes \boldsymbol{\Sigma}^{-1}$ and $(\mathbf{I} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{u}$ respectively. However, there is no explicit formula for $\widehat{\boldsymbol{\Sigma}}$ when $\boldsymbol{\Sigma}$ is not diagonal. Instead, we use the method of Stiratelli, Laird & Ware (1984). Following Stiratelli's method, we assume a diffuse prior

for $\boldsymbol{\beta}$ in addition to the $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ prior for $\mathbf{u}_i$. We can then estimate $\boldsymbol{\Sigma}$ by an EM type estimate

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}(\mathbf{u}_i\mathbf{u}_i^T|\mathbf{Y}) = \frac{1}{m}\sum_{i=1}^{m} \left[\mathrm{E}(\mathbf{u}_i|\mathbf{Y})\mathrm{E}(\mathbf{u}_i^T|\mathbf{Y}) + \mathrm{Cov}(\mathbf{u}_i|\mathbf{Y})\right]. \tag{3.14}$$

Stiratelli suggested approximating the posterior distribution of $(\boldsymbol{\beta}, \mathbf{u})$ given $\mathbf{Y}$ by a normal distribution that has the same mode and curvature as the true posterior. In other words, we can replace the posterior mean $\mathrm{E}(\mathbf{u}_i|\mathbf{Y})$ in (3.14) by the posterior mode $\widehat{\mathbf{u}}_i$ which can be obtained using (3.10). The conditional covariance matrix $\mathrm{Cov}(\mathbf{u}_i|\mathbf{Y})$ in (3.14) can be approximated by inverting the Hessian matrix of the log posterior density of $(\boldsymbol{\beta}, \mathbf{u})$ given $\mathbf{Y}$ which is equivalent to the Hessian matrix of the complete data log-likelihood because of the diffuse prior assumption for $\boldsymbol{\beta}$. By iterating (3.10) and (3.14), we obtain estimates of $\boldsymbol{\beta}, \mathbf{u}$ and $\boldsymbol{\Sigma}$.

3.2 Numerical results

The result of fitting the random effects models can also be found in Table 1 and Table 2 respectively for the AR(1) and AR(2) models. For the random intercept models, the ML and REML estimates are very similar so we report only the former. The Gaussian quadrature estimates GQ_4 ($\kappa = 4$) and the ML estimates are quite similar too. As expected, the standard errors are inflated relative to the model without random intercept but the conclusion is qualitatively the same, namely, reduced heroin use is associated with increased methadone

dose and increased duration in treatment. The estimate of the variance component σ^2 for the random intercept model using ML is 1.244(S.E. = 0.214) for the AR(1) model and 1.042(0.197) for the AR(2) model. In addition to the random intercept model, we also fit models with random intercept β_o and random coefficient for dose β_d and obtain much the same conclusion.

Chapter 4

Mixture model with 2 or more latent groups

4.1 The models

As different patient groups may react differently to methadone treatment, we consider the mixture model. Under this model, we assume that there are two or more groups of patients who respond differently to methadone treatment. Suppose there are G latent groups and each patient has a probability π_k of coming from group $k, k = 1, \ldots, G$. If patient i belongs to group k, then

$$\text{logit}[\Pr(Y_{it} = 1|Y_{i,t-1})] = \eta_{itk} = \beta_{ok} + \beta_{dk} d_{it} + \beta_{tk} \log t + \beta_{p1k} Y_{i,t-1} \qquad (4.1)$$

for the AR(1) model. This is essentially a discrete random effects model such that $\boldsymbol{\beta} = \boldsymbol{\beta}_k = (\beta_{ok}, \beta_{dk}, \beta_{tk}, \beta_{p1k})^T$ with probability π_k. This model has a more tractable likelihood function

$$\prod_{i=1}^{m} \Pr(Y_{i1}, \ldots, Y_{in_i}) = \prod_{i=1}^{m} \Big[\sum_{k=1}^{G} \pi_k \Big(\prod_{t=1}^{n_i} \frac{e^{Y_{it}\eta_{itk}}}{1+e^{\eta_{itk}}} \Big) \Big] \tag{4.2}$$

which does not involve integration. To estimate the parameters, it is convenient to use the EM algorithm. Specifically, we define $W_{ik} = 1$ if patient i belongs to group k and $W_{ik} = 0$ otherwise. Note that W_{ik} are unobserved since the group membership of each patient is unknown. The log-likelihood function based on the so-called 'complete' data $(\mathbf{Y}, \mathbf{W})$ is

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{W}) \;=\; & \sum_{k=1}^{G} \log \pi_k \sum_{i=1}^{m} W_{ik} + \\ & \sum_{k=1}^{G} \sum_{i=1}^{m} \sum_{t=1}^{n_i} W_{ik} Y_{it} \eta_{itk} - \sum_{k=1}^{G} \sum_{i=1}^{m} \sum_{t=1}^{n_i} W_{ik} \log(1 + e^{\eta_{itk}}) \end{aligned} \tag{4.3}$$

where $\boldsymbol{\theta}$ denotes the parameter vector $(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_G^T, \pi_1, \ldots, \pi_{G-1})^T$. At the E-step of the EM algorithm, we estimate the unknown W_{ik} by the conditional expectation

$$\widehat{W}_{ik} = \mathrm{E}(W_{ik}|\mathbf{Y}) = \frac{\pi_k \prod_{t=1}^{n_i} e^{Y_{it}\eta_{itk}}/(1+e^{\eta_{itk}})}{\sum_{k=1}^{G} \pi_k \big[\prod_{t=1}^{n_i} e^{Y_{it}\eta_{itk}}/(1+e^{\eta_{itk}}) \big]} \tag{4.4}$$

evaluated at the current parameter estimates. At the M-step, we replace W_{ik} by $\widehat{W}_{ik}$ in (4.3) and maximize the resulting expression to obtain the updated estimates of the parameters. The procedures are iterated until convergence.

As mentioned earlier, a finite mixture model is just a discrete random effects model. Thus the fitted 2 or 3 group mixture models are actually maximum likelihood estimates (MLE) of the random effects model when the random effect distributions are assumed to have 2 or 3 mass points. Without making any distributional assumption, it is well known that the nonparametric MLE

of the random effect distribution is a discrete distribution with a finite number of mass points (Laird, 1978, Lindsay, 1983). Thus we can obtain the nonparametric MLE via fitting a finite mixture model by increasing the number of groups until the likelihood ceases to increase.

There is also a connection between Gaussian quadrature and finite mixture maximum likelihood. The number of quadrature points used in Gaussian quadrature plays the same role as the number of components in the finite mixture (Aitkin, 1996).

4.2 Numerical results

We fit the mixture model with different number of groups to the data and the results are given in Table 5. There is almost no increase in likelihood beyond 3 groups which explains why the AIC criterion picks a 3 group mixture model with group specific intercept and dose coefficient as the best model. The decision of keeping the time and autoregression coefficients fixed across the different groups is also based on the AIC criterion. Since the time coefficient is negative and significant, we conclude that there is an association between reduced heroin use and increased duration of treatment. The dose coefficient is not significant in group 1 ($\pi_1 = 0.30$) and group 3 ($\pi_3 = 0.24$) suggesting that heroin use is independent of daily methadone dose for patients in these groups. To gain more knowledge about the makeup of these groups, we look at $\widehat{W}_{ik}$, the estimated probability that subject i belongs to group k. From Table 6, we can see that subjects who are undoubtedly in group 1 ($\widehat{W}_{i1} \geq 0.9$) are those with no or very few positive results for heroin. Thus subjects in group 1

are those who have ceased heroin use as a result of treatment. Individuals with $\widehat{W}_{i3} \geq 0.9$ are those with a persistently large number of heroin positive screens. These subjects respond poorly to treatment, with continued heroin use regardless of the methadone dose received. It is only in group 2 ($\pi_2 = 0.46$) that methadone dose is significant. Subjects in this group respond to treatment in a dose-dependent fashion with reduced heroin use at high methadone dose.

4.3 Discussion

We found that the best fit statistical model is a 3 group mixture model. The first group comprise about 30% of the sample, and respond well to treatment, with cessation of heroin use. Methadone dose is not predictive of heroin positive urines in this group, consistent with clinical experience that there is a proportion of patients who have an excellent response to treatment, and can be satisfactorily maintained on quite modest doses of methadone. The second group, 46% of the sample, tend to use heroin when on lower methadone doses, but when maintained on high doses do quite well. Again, this fits well with clinical experience. Finally, there is a group, 24% in this study, who tend to continue to use heroin, regardless of methadone dose. It is well recognised that all treatment programs have to confront the problem of treatment failures, although the proportion of such failures probably varies according to the quality of other aspects of treatment (Ball & Ross, 1991). The current study is useful in identifying that methadone dose is of critical importance in influencing heroin use in around half the patients in treatment.

Chapter 5

Bivariate binary model

5.1 The models

As multiple drug use is common among the methadone clinic patients, we extend the model to bivariate binary data so that we are able to study the use of heroin, benzodiazepines and their interaction simultaneously. Let $Y_{itj}(i = 1, \ldots, m;\ t = 1, \ldots, n_i;\ j = h$ for heroin, $j = b$ for benzodiazepines) denote the observed outcome of drug j for the i-th patient at time t. A conditional approach is to model simultaneously the conditional probabilities $P_{it}(1, \cdot) = \Pr(Y_{ith} = 1 \mid Y_{i,t-1,h}, Y_{i,t-1,b})$, $P_{it}(\cdot, 1) = \Pr(Y_{itb} = 1 \mid Y_{i,t-1,h}, Y_{i,t-1,b})$, $P_{it}(u, v) = \Pr(Y_{ith} = u, Y_{itb} = v \mid Y_{i,t-1,h}, Y_{i,t-1,b})$ as well as the odds ratio

$$\gamma_{it} = \frac{P_{it}(1,1)\ P_{it}(0,0)}{P_{it}(1,0)\ P_{it}(0,1)} \tag{5.1}$$

such that

$$\text{logit}[P_{it}(1, \cdot)] \ = \eta_{ith} = \ \beta_{oh} + \beta_{dh} d_{it} + \beta_{th} \log t + \beta_{ph,h} Y_{i,t-1,h} +$$

$$\beta_{ph,b}Y_{i,t-1,b} + \beta_{ph,h*b}Y_{i,t-1,h}Y_{i,t-1,b}, \tag{5.2}$$

$$\begin{aligned}\text{logit}[P_{it}(\cdot,1)] \quad = \eta_{itb} = \quad & \beta_{ob} + \beta_{db}d_{it} + \beta_{tb}\log t + \beta_{pb,h}Y_{i,t-1,h} + \\ & \beta_{pb,b}Y_{i,t-1,b} + \beta_{pb,h*b}Y_{i,t-1,h}Y_{i,t-1,b}\end{aligned} \tag{5.3}$$

and the log odds ratio

$$\begin{aligned}\log[\gamma_{it}] \quad = \eta_{itr} = \quad & \beta_{or} + \beta_{dr}d_{it} + \beta_{tr}\log t + \beta_{pr,h}Y_{i,t-1,h} + \\ & \beta_{pr,b}Y_{i,t-1,b} + \beta_{pr,h*b}Y_{i,t-1,h}Y_{i,t-1,b}.\end{aligned} \tag{5.4}$$

Note that for simplicity, we only consider the AR(1) model.

5.1.1 Bivariate autoregressive model

Let $\boldsymbol{\theta}_j$ denote the $q_j \times 1$ vector of $(\beta_{oj}, \beta_{dj}, \beta_{tj}, \beta_{pj,h}, \beta_{pj,b}, \beta_{pj,h*b})$, $j = h, b$ or r. Then the joint conditional probability $P_{it}(1,1)$ can be expressed (Fleiss, 1981, P.68) in terms of $P_{it}(1,\cdot), P_{it}(\cdot,1)$ and the odds ratio γ_{it} as follows

$$P_{it}(1,1) = \frac{(\gamma_{it}-1)\left[P_{it}(1,\cdot)+P_{it}(\cdot,1)\right] + 1 - \delta_{it}^{\frac{1}{2}}}{2(\gamma_{it}-1)} \tag{5.5}$$

where

$$\begin{aligned}\delta_{it} \quad = \quad & 1 + (\gamma_{it}-1)\big\{\gamma_{it}[P_{it}(1,\cdot) - P_{it}(\cdot,1)]^2 - \\ & [P_{it}(1,\cdot) + P_{it}(\cdot,1)]^2 + 2[P_{it}(1,\cdot) + P_{it}(\cdot,1)]\big\}.\end{aligned} \tag{5.6}$$

We also have $P_{it}(1,0) = P_{it}(1,\cdot) - P_{it}(1,1)$, $P_{it}(0,1) = P_{it}(\cdot,1) - P_{it}(1,1)$ and $P_{it}(0,0) = 1 - P_{it}(1,\cdot) - P_{it}(\cdot,1) + P_{it}(1,1)$. The likelihood function is

$$\prod_{i=1}^{m}\prod_{t=1}^{n_i} \Pr(Y_{ith}, Y_{itb} \,|Y_{i,t-1,h}, Y_{i,t-1,b}) = \prod_{i=1}^{m}\prod_{t=1}^{n_i} P_{it}(Y_{ith}, Y_{itb}) \tag{5.7}$$

and the log-likelihood function is

$$\begin{aligned} l \;=\; \sum_{i=1}^{m}\sum_{t=1}^{n_i} \big[\; & Y_{ith}Y_{itb}\,\log P_{it}(1,1) \;+\; Y_{ith}(1-Y_{itb})\,\log P_{it}(1,0) \;+ \\ & (1-Y_{ith})Y_{itb}\,\log P_{it}(0,1) \;+ \\ & (1-Y_{ith})(1-Y_{itb})\,\log P_{it}(0,0)\big]. \end{aligned} \tag{5.8}$$

The Newton-Raphson method is used to obtain the maximum likelihood estimates. The first and second order derivatives of the log-likelihood function required in the Newton-Raphson procedure are quite tedious and are given in the Appendix. Zeger & Liang (1991, equation 2.3) assumed (5.2) and (5.3) above but not (5.4). As the model is not completely specified without (5.4), they resorted to a GEE approach . Liang & Zeger (1989) modelled in terms of $\Pr(Y_{ith}\,|Y_{itb}$, past) and $\Pr(Y_{itb}\,|Y_{ith}$, past) but since the likelihood function is intractable, they resorted to a pseudolikelihood estimation procedure.

5.1.2 Bivariate mixture model

Under this model, we assume that there are two groups of patients who react differently to methadone treatment. Suppose each patient has a probability π_k of coming from group $k, k = 1, 2$. If patient i belongs to group k, then

$$\text{logit}[\Pr(Y_{ith} = 1|Y_{i,t-1,h}, Y_{i,t-1,b})] = \text{logit}[P_{itk}(1,\cdot)] = \eta_{ithk} = \beta_{ohk} + \tag{5.9}$$

$$\beta_{dhk}d_{it} + \beta_{thk}\log t + \beta_{ph,hk}Y_{i,t-1,h} + \beta_{ph,bk}Y_{i,t-1,b} + \beta_{ph,h*b,k}Y_{i,t-1,h}Y_{i,t-1,b},$$

$$\begin{aligned}\text{logit}[\Pr(Y_{itb} = 1|Y_{i,t-1,h}, Y_{i,t-1,b})] = \text{logit}[P_{itk}(\cdot,1)] = \eta_{itbk} = \beta_{obk} + \\ \beta_{dbk}d_{it} + \beta_{tbk}\log t + \beta_{pb,hk}Y_{i,t-1,h} + \beta_{pb,bk}Y_{i,t-1,b} + \beta_{pb,h*b,k}Y_{i,t-1,h}Y_{i,t-1,b}\end{aligned} \quad (5.10)$$

and we assume that the odds ratio is constant across the two groups as in (5.4). The joint probability $P_{itk}(1,1)$ for group k is defined in terms of the marginal probabilities $P_{itk}(1,\cdot)$ and $P_{itk}(\cdot,1)$ and the odds ratio γ_{it} as in (5.5) and (5.6). This model has a tractable likelihood function which does not involve integration

$$\begin{aligned}\prod_{i=1}^{m}\Pr(Y_{i1},\ldots,Y_{in_i}) &= \prod_{i=1}^{m}\Big[\sum_{k=1}^{2}\pi_k\Big(\prod_{t=1}^{n_i}P_{itk}(1,1)^{Y_{ith}Y_{itb}}\ P_{itk}(1,0)^{Y_{ith}(1-Y_{itb})} \\ &\quad P_{itk}(0,1)^{(1-Y_{ith})Y_{itb}}\ P_{itk}(0,0)^{(1-Y_{ith})(1-Y_{itb})}\Big)\Big] \quad (5.11)\end{aligned}$$

where $P_{itk}(1,0) = P_{itk}(1,\cdot) - P_{itk}(1,1)$, $P_{itk}(0,1) = P_{itk}(\cdot,1) - P_{itk}(1,1)$ and $P_{itk}(0,0) = 1 - P_{itk}(1,\cdot) - P_{itk}(\cdot,1) + P_{itk}(1,1)$. To estimate the parameters, it is convenient to use the EM algorithm. Specifically, we define $W_{ik} = 1$ if patient i belongs to group k and $W_{ik} = 0$ otherwise. Note that W_{ik} is unobserved since the group membership of such patient is unknown. The log-likelihood function based on the so-called 'complete' data $(\mathbf{Y}, \mathbf{W})$ is

$$\begin{aligned}l(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{W}) &= \sum_{k=1}^{2}\Big(\log\pi_k\sum_{i=1}^{m}W_{ik}\Big) + \quad (5.12) \\ &\quad \sum_{k=1}^{2}\sum_{i=1}^{m}\sum_{t=1}^{n_i}W_{ik}\Big[Y_{ith}Y_{itb}\log P_{itk}(1,1) + Y_{ith}(1-Y_{itb})\log P_{itk}(1,0) + \\ &\quad (1-Y_{ith})Y_{itb}\log P_{itk}(0,1) + (1-Y_{ith})(1-Y_{itb})\log P_{itk}(0,0)\Big]\end{aligned}$$

where $\boldsymbol{\theta}$ denotes the parameter vector $(\boldsymbol{\theta}_{h1}^T, \boldsymbol{\theta}_{h2}^T, \boldsymbol{\theta}_{b1}^T, \boldsymbol{\theta}_{b2}^T, \boldsymbol{\theta}_r^T, \pi_1)^T$. At the E-step of the EM algorithm, we estimate the unknown W_{ik} by the conditional expectation

$$\widehat{W}_{ik} = \mathrm{E}(W_{ik}|\mathbf{Y}) = \frac{\pi_k \prod_{t=1}^{n_i} P_{itk}(1,1)^{Y_{ith}Y_{itb}} \; P_{itk}(1,0)^{Y_{ith}(1-Y_{itb})} \; P_{itk}(0,1)^{(1-Y_{ith})Y_{itb}} \; P_{itk}(0,0)^{(1-Y_{ith})(1-Y_{itb})}}{\sum_{k=1}^{2} \pi_k \left[\prod_{t=1}^{n_i} P_{itk}(1,1)^{Y_{ith}Y_{itb}} P_{itk}(1,0)^{Y_{ith}(1-Y_{itb})} P_{itk}(0,1)^{(1-Y_{ith})Y_{itb}} P_{itk}(0,0)^{(1-Y_{ith})(1-Y_{itb})} \right]} \tag{5.13}$$

evaluated at the current parameter estimates. At the M-step, we replace W_{ik} by $\widehat{W}_{ik}$ in (5.12) and maximize the resulting expression to obtain the updated estimates of the parameters using Newton-Raphson method. The procedure is iterated until convergence.

5.2 Numerical results

The results of fitting the full model defined by (5.2), (5.3) and (5.4) above are given in Table 7 as model 1. By discarding the nonsignificant coefficients, we arrive at model 2. It is interesting to note that for the odds ratio regression, $\beta_{pr,h} \simeq \beta_{pr,b}$ and $\beta_{pr,h} + \beta_{pr,b} + \beta_{pr,h*b} \simeq 0$ which suggest that the odds ratio γ_{it} depends on the previous outcomes only through whether $Y_{i,t-1,h} = Y_{i,t-1,b}$ or not. The p-values of the z-test for testing the two constraints separately are 0.9124 and 0.7642 respectively which suggest that the two constraints do hold. If we define the *concordance indicator* $C_{i,t-1}$ to be 1 if $Y_{i,t-1,h} = Y_{i,t-1,b}$ and 0 otherwise and assume that

$$\log[\gamma_{it}] = \eta_{itr} = \beta_{or} + \beta_{dr} d_{it} + \beta_{cr} C_{i,t-1}, \tag{5.14}$$

we have model 3 in Table 7. We can see that the AIC value under model 3 is much improved. The likelihood ratio test for testing the concordance model against the previous model is nonsignificant which again suggests that the concordance model is to be preferred.

We can see from Table 7 that both the dose and time effect for both drugs are significant. Our conclusion is that reduced drug use is associated with increased duration in treatment. However, while reduced heroin use is associated with increased methadone dose, increased benzodiazepines use is also associated with increased methadone dose. This suggests that in contrast to the dramatic pharmacological effect of methadone in reducing heroin use and psychological dependence on heroin, methadone maintenance does not suppress non-opioid drug use. There is a strong positive association between the present and the previous outcomes of the same drugs. In fact, some patients in treatment tend to use drugs continuously. For the odds ratio, it seems to depend on whether the previous outcomes for heroin and benzodiazepines are concordant or discordant. This suggests that the strength of the present association between the two drugs depends on the strength of their association last week.

The results of fitting various bivariate binary mixture models with two latent groups are given in Table 8. We start with the same covariates as model 3 and fit a two-group mixture model with group-specific regression coefficients to result in model 4. By discarding nonsignificant covariates and setting the regression coefficients to be constant across the two groups if the group-specific estimates are similar, we have model 5 and 6. Based on AIC, we choose

model 5. We can easily classify most of the patients into one of the two groups as $\widehat{W}_{i1}$ is very close to 0 or 1 for most i. By classifying patient i to group k if $\widehat{W}_{ik} > \frac{1}{2}$, we can see that the mixture model tends to divide patients into light drug user group (group 1) and heavy drug user group (group 2). Specifically, group 1 consists of 75 patients submitting 6.9% positive screen for morphine and 2.6% positive screen for benzodiazepines from a total of 1667 screens and group 2 consists of 61 patients submitting 29.4% positive screen for morphine and 33.6% positive screen for benzodiazepines from a total of 1205 screens. The conclusions for dose and time effect are similar to those based on a non-mixture model. Again, there is a strong positive association between the present and the previous outcomes of the same drugs. It is interesting to note that the association between the present outcome of one drug and the previous outcome of the other drug is significant and negative only in group 2. This suggests that patients in group 2 are not heavy user of both drugs: they only use one particular drug heavily.

Chapter 6

Probit-linear mixed model using MLE

In this Chapter, we will consider the use of probit link. McCulloch (1994) pointed out several advantages of using the probit link instead of the customary logit link. For example, the probit link function is preserved when modelling the marginal distribution. Moreover, by viewing a probit-normal model as a threshold model that results from dichotomizing some unobserved continuous observations from a Gaussian mixed model, it becomes feasible to use the EM algorithm to find the maximum likelihood estimates.

6.1 The model

Let $W_1, \ldots, W_n$ denote the observed binary variables. Following McCulloch (1994), we assume that the probabilities $P_i = \Pr(W_i = 1)$ are probit-linear. In

matrix form,

$$\Phi^{-1}(\mathbf{P}) = \left(\Phi^{-1}(P_1), \ldots, \Phi^{-1}(P_n)\right)^T = \mathbf{X}\boldsymbol{\beta} + \sum_{r=1}^{R} \mathbf{Z}_r \mathbf{u}_r \tag{6.1}$$

where $\mathbf{X}$ is a $n \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{u}_r$ is a $q_r k_r \times 1$ vector of random effects with corresponding $n \times q_r k_r$ design matrix $\mathbf{Z}_r$. We assume that $\mathbf{u}_1, \ldots, \mathbf{u}_R$ are independent. For each $\mathbf{u}_r$, we have

$$\mathbf{u}_r = \begin{pmatrix} \mathbf{u}_{r1} \\ \vdots \\ \mathbf{u}_{rq_r} \end{pmatrix} \sim \mathcal{N}_{q_r \times k_r} \left\{ \mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma}_r & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{\Sigma}_r \end{pmatrix} \right\}. \tag{6.2}$$

In other words, $\mathbf{u}_r$ is made up of q_r i.i.d. random vectors of dimension k_r each.

For the purpose of estimation, it is useful to view the above probit-linear mixed model as a threshold model that results from dichotomizing the observations from a Gaussian mixed model. In other words, $W_i = I_{(Y_i > 0)}$ and

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{r=1}^{R} \mathbf{Z}_r \mathbf{u}_r + \boldsymbol{\varepsilon} \tag{6.3}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ independently of the $\mathbf{u}_r$.

The above model allows correlated random effects and so is an extension of the model proposed by McCulloch (1994) which considers only independent random effects. The extended model is quite general and includes the following useful models as special cases.

1. McCulloch's model

   This model is obtained by setting $k_r \equiv 1$.

2. Random coefficients regression model for clustered binary data

   Let W_{it}, $i = 1, \ldots, m$ and $t = 1, \ldots, n_i$ represents the t-th observation from the i-th cluster. We assume that $W_{it} = I_{(Y_{it}>0)}$ where

   $$Y_{it} = \mathbf{x}_{it}\mathbf{b}_i + \varepsilon_{it},$$

   $\mathbf{x}_{it}$ denotes a $1 \times p$ vector of covariates and $\mathbf{b}_i$ denotes the corresponding cluster-specific $p \times 1$ vector of regression coefficients. If we assume that the m cluster-specific vectors of regression coefficients, $\mathbf{b}_1, \ldots, \mathbf{b}_m$ are i.i.d. $\mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, we can rewrite $\mathbf{b}_i$ as $\boldsymbol{\beta} + \mathbf{u}_i$ and we get

   $$Y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{x}_{it}\mathbf{u}_i + \varepsilon_{it}.$$

   Putting it in vector form, we have

   $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}\mathbf{u} + \boldsymbol{\varepsilon}.$$

   This model corresponds to (6.3) with $R = 1$, $\mathbf{Z}_1 = \mathbf{X}$, $q_1 = m$ and $k_1 = p$.

3. Random intercept model for bivariate clustered binary data

   Let W_{1it} and W_{2it} denote the two binary variables observed for unit t of cluster i. We assume that $W_{kit} = I_{(Y_{kit}>0)}$ where

   $$Y_{1it} = u_{1i} + x_{it}\beta_1 + \varepsilon_{1it},$$
   $$Y_{2it} = u_{2i} + x_{it}\beta_2 + \varepsilon_{2it}.$$

   In other words, we have a random intercept model for each variable and the random intercepts $(u_{1i}, u_{2i})^T$ are i.i.d. according to a bivariate normal distribution with zero mean. This model is a special case of (6.3) with $R = 1, q_1 = m$ and $k_1 = 2$.

4. Models involving crossed design of correlated random effects

   Two complicated models of crossed design are illustrated in Section 6.2 through the study of the famous salamander data.

6.1.1 Maximum likelihood estimation

In this Section, we describe the maximum likelihood estimation of the fixed effects $\boldsymbol{\beta}$ and the variance components $\boldsymbol{\Sigma}_r, r = 1, \ldots, R$, by the EM algorithm. To apply the EM algorithm, we treat the complete data as $\mathbf{Y}, \mathbf{u}_1, \ldots, \mathbf{u}_R$ and regard the observed data $\mathbf{W}$ as the incomplete data. The EM algorithm is particularly suited for the probit-normal model given by (6.2) and (6.3) as closed form formulae for the complete data MLE exist and they are

$$\widehat{\boldsymbol{\beta}}_c = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T(\mathbf{Y} - \sum_{r=1}^{R}\mathbf{Z}_r\mathbf{u}_r) \tag{6.4}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{rc} = (\sum_{j=1}^{q_r}\mathbf{u}_{rj}\mathbf{u}_{rj}^T) \,/\, q_r. \tag{6.5}$$

Instead of (6.4), McCulloch (1994) used $\widehat{\boldsymbol{\beta}}(\mathbf{Y}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}$ where $\mathbf{V}$ is the covariance matrix of $\mathbf{Y}$. Note that $\widehat{\boldsymbol{\beta}}(\mathbf{Y})$ is the MLE for $\boldsymbol{\beta}$ based on $\mathbf{Y}$ alone rather than based on the complete data $\mathbf{Y}, \mathbf{u}_1, \ldots, \mathbf{u}_R$. McCulloch's procedure is closely related to version 1 of the ECME algorithm proposed by Liu & Rubin (1994, P.641).

As the complete data are distributed according to an exponential family, the EM procedure can be simplified considerably (Little & Rubin, 1987, §7.6) as follows. Given the current estimates $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\Sigma}_r^{(k)}$, $r = 1, \ldots, R$, we update our parameter estimates by replacing $\mathbf{Y}$, $\mathbf{u}_r$ and $\mathbf{u}_{rj}\mathbf{u}_{rj}^T$ in (6.4) and (6.5) by their conditional expectation evaluated at the current parameter estimates to get

$$\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T(\,\mathrm{E}[\mathbf{Y}|\mathbf{W}]\; -\; \sum_{r=1}^{R}\mathbf{Z}_r\mathrm{E}[\mathbf{u}_r|\mathbf{W}]\,) \tag{6.6}$$

and

$$\boldsymbol{\Sigma}_r^{(k+1)} = (\,\sum_{j=1}^{q_r}\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^T|\mathbf{W}]\,)\;/\;q_r. \tag{6.7}$$

The steps are then iterated until convergence is achieved. Now

$$\mathrm{E}[\mathbf{u}_r|\mathbf{W}] = \mathrm{E}[\mathrm{E}[\mathbf{u}_r|\mathbf{Y}]|\mathbf{W}], \tag{6.8}$$

$$\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^T|\mathbf{W}] = \mathrm{E}[\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^T|\mathbf{Y}]|\mathbf{W}]. \tag{6.9}$$

To find the inner expectations, we use the fact that the joint distribution of $\mathbf{Y}, \mathbf{u}_1, \ldots, \mathbf{u}_R$ are

$$\begin{pmatrix} Y \\ u_1 \\ \vdots \\ u_R \end{pmatrix} \sim \mathcal{N}_{n+Q}\left\{ \begin{pmatrix} X\beta \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} V & Z_1(I_{q_1}\otimes\Sigma_1) & \cdots & \cdots & Z_R(I_{q_R}\otimes\Sigma_R) \\ (I_{q_1}\otimes\Sigma_1)Z_1^T & (I_{q_1}\otimes\Sigma_1) & 0 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ (I_{q_R}\otimes\Sigma_R)Z_R^T & 0 & \cdots & 0 & (I_{q_R}\otimes\Sigma_R) \end{pmatrix} \right\}$$

where

$$\mathbf{V} = \mathbf{I}_n \; + \; \sum_{r=1}^{R}\mathbf{Z}_r(\mathbf{I}_{q_r}\otimes\boldsymbol{\Sigma}_r)\mathbf{Z}_r^T \tag{6.10}$$

and $Q = \sum_{r=1}^{R} q_r k_r$. It follows that

$$\mathrm{E}[\mathbf{u}_r|\mathbf{Y}] = (\mathbf{I}_{q_r} \otimes \boldsymbol{\Sigma}_r)\mathbf{Z}_r^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{6.11}$$

and

$$\begin{aligned}\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^T|\mathbf{Y}] &= \boldsymbol{\Sigma}_r\mathbf{Z}_{rj}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{Z}_{rj}\boldsymbol{\Sigma}_r + \\ &\quad \boldsymbol{\Sigma}_r - \boldsymbol{\Sigma}_r\mathbf{Z}_{rj}^T\mathbf{V}^{-1}\mathbf{Z}_{rj}\boldsymbol{\Sigma}_r \end{aligned} \tag{6.12}$$

where $\mathbf{Z}_{rj}$ is the $n \times k_r$ submatrix of $\mathbf{Z}_r = (\mathbf{Z}_{r1}, \ldots, \mathbf{Z}_{rq_r})$ that corresponds to $\mathbf{u}_{rj}$. Substituting (6.11) and (6.12) into (6.8) and (6.9), we obtain

$$\mathrm{E}[\mathbf{u}_r|\mathbf{W}] = (\mathbf{I}_{q_r} \otimes \boldsymbol{\Sigma}_r)\mathbf{Z}_r^T\mathbf{V}^{-1}(\mathrm{E}[\mathbf{Y}|\mathbf{W}] - \mathbf{X}\boldsymbol{\beta}) \tag{6.13}$$

and

$$\begin{aligned}\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^T|\mathbf{W}] &= \boldsymbol{\Sigma}_r\mathbf{Z}_{rj}^T\mathbf{V}^{-1}\{\mathrm{E}[\mathbf{Y}|\mathbf{W}]\mathrm{E}[\mathbf{Y}|\mathbf{W}]^T + \mathrm{Cov}[\mathbf{Y}|\mathbf{W}] - \\ &\quad \mathrm{E}[\mathbf{Y}|\mathbf{W}](\mathbf{X}\boldsymbol{\beta})^T - (\mathbf{X}\boldsymbol{\beta})\mathrm{E}[\mathbf{Y}|\mathbf{W}]^T + (\mathbf{X}\boldsymbol{\beta})(\mathbf{X}\boldsymbol{\beta})^T\} \\ &\quad \mathbf{V}^{-1}\mathbf{Z}_{rj}\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_r - \boldsymbol{\Sigma}_r\mathbf{Z}_{rj}^T\mathbf{V}^{-1}\mathbf{Z}_{rj}\boldsymbol{\Sigma}_r. \end{aligned} \tag{6.14}$$

We can see from (6.13) and (6.14) that $\mathrm{E}[\mathbf{u}_r|\mathbf{W}]$ and $\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^T|\mathbf{W}]$ can be expressed entirely in terms of the conditional mean and the conditional covariance matrix of $\mathbf{Y}$ given $\mathbf{W}$. Thus $\mathrm{E}[\mathbf{Y}|\mathbf{W}]$ and $\mathrm{Cov}[\mathbf{Y}|\mathbf{W}]$ are all that we need to carry out the EM iterations in (6.6) and (6.7).

6.1.2 The Monte Carlo EM algorithm

Unfortunately, E[$\mathbf{Y}|\mathbf{W}$] and Cov[$\mathbf{Y}|\mathbf{W}$] which are required for the EM algorithm cannot be expressed in closed form. In such circumstances, Wei & Tanner (1990) and McCulloch (1994) suggest the so-called Monte Carlo EM algorithm where E[$\mathbf{Y}|\mathbf{W}$] and Cov[$\mathbf{Y}|\mathbf{W}$] are approximated by simulations. Since it is difficult to simulate directly from the conditional distribution of $\mathbf{Y}$ given $\mathbf{W}$, we use the method of Gibbs sampling (Smith & Roberts, 1993). The details are given as follows.

Let us first denote the vector obtained by taking the diagonal and lower triangular elements of $\mathbf{\Sigma}_r$ by vec$\mathbf{\Sigma}_r$ and the $P \times 1$ vector of all parameters $(\boldsymbol{\beta}^T, \text{vec}\mathbf{\Sigma}_1{}^T, \ldots, \text{vec}\mathbf{\Sigma}_R{}^T)^T$ by $\boldsymbol{\theta}$ where $P = p + \sum_{r=1}^{R} q_r(q_r+1)/2$. Starting with some initial values $\mathbf{Y}^{(0)} = (Y_1^{(0)}, \ldots, Y_n^{(0)})^T$ consistent in signs with the observed data $\mathbf{W}$, we proceed to generate $\mathbf{Y}^{(b)} = (Y_1^{(b)}, \ldots, Y_n^{(b)})^T$, $b = 1, 2, \ldots$ sequentially in the following manner.

Given $\mathbf{Y}^{(b)}$ and the current estimates $\boldsymbol{\theta}^{(k)}$, we simulate

$$
\begin{aligned}
&Y_1^{(b+1)} \text{ from } f(y_1|y_2^{(b)}, y_3^{(b)}, \ldots, y_n^{(b)}, \mathbf{W}; \boldsymbol{\theta}^{(k)}) \\
&\vdots \\
&Y_i^{(b+1)} \text{ from } f(y_i|y_1^{(b+1)}, y_2^{(b+1)}, \ldots, y_{i-1}^{(b+1)}, y_{i+1}^{(b)}, \ldots, y_n^{(b)}, \mathbf{W}; \boldsymbol{\theta}^{(k)}) \quad (6.15) \\
&\vdots \\
&Y_n^{(b+1)} \text{ from } f(y_n|y_1^{(b+1)}, y_2^{(b+1)}, \ldots, y_{n-1}^{(b+1)}, \mathbf{W}; \boldsymbol{\theta}^{(k)}).
\end{aligned}
$$

To carry out the above simulations, we use the fact that

$$f(y_i|y_j, j \neq i; w_1, \ldots, w_n) = f(y_i|y_j, j \neq i; w_i).$$

Since $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is distributed as $\mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ with $\mathbf{V}$ given by (6.10), it follows that $f(y_i|y_j, j \neq i)$ is also normal. Thus $f(y_i|y_j, j \neq i; w_i)$ is a truncated (above 0 if $w_i = 1$; below 0 if $w_i = 0$) normal distribution. We use the fast acceptance-rejection method of Marsaglia (1964) to simulate from such a truncated normal distribution. In this way, we can carry out the Gibbs sampling cycles (6.15) to result in $\mathbf{Y}^{(b)} = (Y_1^{(b)}, \ldots, Y_n^{(b)})^T$, $b \geq 1$. According to the theory of Gibbs sampling, $\{\mathbf{Y}^{(b)}\}$ is a Markov chain whose stationary distribution is $f(\mathbf{y}|\mathbf{w})$ which we wish to sample from. In our analysis, we discard the first $T = 200$ elements of the Markov chain $\{\mathbf{Y}^{(b)}\}$ as transient values and treat the following $B = 1000$ elements of $\{\mathbf{Y}^{(b)}\}$ as realizations from the conditional distribution of $\mathbf{Y}$ given $\mathbf{W}$. We use their sample mean and sample covariance matrix to approximate $\mathrm{E}[\mathbf{Y}|\mathbf{W}]$ and $\mathrm{Cov}[\mathbf{Y}|\mathbf{W}]$. To be precise, we have

$$\widehat{\mathrm{E}}[\mathbf{Y}|\mathbf{W}] = \overline{\mathbf{Y}} = \frac{1}{B}\sum_{b=T+1}^{T+B} \mathbf{Y}^{(b)} \tag{6.16}$$

and

$$\widehat{\mathrm{Cov}}[\mathbf{Y}|\mathbf{W}] = \frac{1}{B}\sum_{b=T+1}^{T+B} (\mathbf{Y}^{(b)} - \overline{\mathbf{Y}})(\mathbf{Y}^{(b)} - \overline{\mathbf{Y}})^T. \tag{6.17}$$

Therefore, by substituting (6.13), (6.14), (6.16) and (6.17) back to (6.6) and (6.7), we are able to update the current estimates to $\boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{\Sigma}_r^{(k+1)}$, $r = 1, \ldots, R$ and we continue to iterate until convergence is reached to result in the final estimates $\widehat{\boldsymbol{\theta}}$. Note that we are able to work out the estimates of the random effects as a by-product from (6.13).

6.1.3 Monte Carlo approximation of the observed information matrix

McCulloch (1994) did not report standard errors in his examples but he did mention the supplemented EM (SEM) algorithm of Meng & Rubin (1991) as a possible method. The basic idea of the SEM algorithm is to use the fact that the fraction of missing information is related to the rate of convergence of the EM. By running a sequence of supplementary EM iterations, we can approximate the rate of convergence of the EM algorithm by using finite differences. In this way, we can estimate the increased variability due to missing information which can then be added to the complete data variance-covariance matrix. We do not recommend the SEM algorithm for problems requiring Monte Carlo E-steps. At each iteration of the SEM procedure, we need to consider the P sets of parameter values that result from perturbing the P components of $\boldsymbol{\theta}$ one at a time. For each of these P sets of parameter values, we need to run one iteration of the EM algorithm via Gibbs sampling. In other words, we have to carry out Gibbs sampling P times at each step of the SEM algorithm. This is very time consuming if P is large. For the models we proposed in Section 6.2, $P = 13$ and 21. Thus to run 100 SEM steps, we have to carry out Gibbs sampling 1300 times for model 1 and 2100 times for model 2. Moreover, our experience suggests that a Monte Carlo implementation of the SEM algorithm is numerically unstable, has convergence problem and sometimes leads to negative variance estimates. This is somewhat surprising as no such problems are reported in the literature for the SEM algorithm. A possible explanation is that the finite difference method of approximating the rate of convergence matrix DM (Meng & Rubin, 1991, P.902) is adversely affected by the extra variation due to Monte Carlo sampling. In conclusion, the Monte Carlo SEM algorithm is undesirable in terms of both computing time and numerical stability.

Instead of using the SEM algorithm, we will use simulations to approximate the observed information matrix directly. The details are as follows. Let $l(\boldsymbol{\theta}; \mathbf{W}) = \log f(\mathbf{w}; \boldsymbol{\theta})$ denote the log-likelihood function based on the observed data $\mathbf{W}$. Louis (1982) expressed $l''(\boldsymbol{\theta}; \mathbf{W})$ in terms of certain conditional expectations of the derivatives of the complete data log-likelihood $l(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U})$ given the observed data $\mathbf{W}$. Specifically,

$$\begin{aligned} l''(\boldsymbol{\theta}; \mathbf{W}) &= \mathrm{E}[l''(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U})|\mathbf{W}; \boldsymbol{\theta}] + \mathrm{E}[l'(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U}) l'^{T}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U})|\mathbf{W}; \boldsymbol{\theta}] - \\ &\quad \mathrm{E}[l'(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U})|\mathbf{W}; \boldsymbol{\theta}] \, \mathrm{E}[l'(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U})|\mathbf{W}; \boldsymbol{\theta}]^{T} \end{aligned} \tag{6.18}$$

where $\mathbf{U} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_R^T)^T$. By simulating $(\mathbf{Y}_m, \mathbf{U}_m)$, $m = 1, \ldots, M$ from the conditional distribution of $(\mathbf{Y}, \mathbf{U})$ given $\mathbf{W}$, we can approximate the conditional expectations involved in (6.18) by the corresponding sample means to obtain

$$\begin{aligned} l''_M &= \frac{1}{M} \sum_{m=1}^{M} l''(\widehat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m) + \\ &\quad \frac{1}{M} \sum_{m=1}^{M} l'(\widehat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m) \, l'^{T}(\widehat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m) - \\ &\quad \{\frac{1}{M} \sum_{m=1}^{M} l'(\widehat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m)\} \{\frac{1}{M} \sum_{m=1}^{M} l'(\widehat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m)\}^{T} \end{aligned} \tag{6.19}$$

as a Monte Carlo estimate of $l''(\widehat{\boldsymbol{\theta}}; \mathbf{W})$ and the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$ is estimated by $-l''^{-1}_M$. In our example, M is increased gradually until the estimated standard errors become stable.

We now describe how to simulate from the conditional distribution of $(\mathbf{Y}, \mathbf{U})$ given $\mathbf{W}$. Since

$$f(\mathbf{y}, \mathbf{u}|\mathbf{w}) = f(\mathbf{y}|\mathbf{w})f(\mathbf{u}|\mathbf{y}, \mathbf{w}) = f(\mathbf{y}|\mathbf{w})f(\mathbf{u}|\mathbf{y}), \tag{6.20}$$

we can first simulate $\mathbf{Y}$ from $f(\mathbf{y}|\mathbf{w})$ using the Gibbs sampling technique described in Section 6.1.2 and then simulate $\mathbf{U}$ from $f(\mathbf{u}|\mathbf{y})$ which is normal with mean $\boldsymbol{\mu}_{U|Y}$ and covariance matrix $\boldsymbol{\Sigma}_{U|Y}$ obtainable from standard formulae.

6.1.4 Accounting for Monte Carlo variation

Since Gibbs sampling is used to approximate the various conditional expectations required at the E-step of the algorithm, we need to check whether the Gibbs sampler has converged. While a lot of stopping criteria have been proposed in the literature, they are too microscopic in nature and are not designed with Monte Carlo maximum likelihood estimation in mind. Specifically, the existing criteria are primarily concerned with simulations from one target distribution to approximate an expectation. In contrast, we need to approximate a lot of expectations and the distribution that we wish to sample from changes with each iteration as the parameter estimates are updated. Furthermore, our primary interest is not in the expectations themselves but in the parameter estimates they eventually lead to. In view of the above, we decide to adopt a more macroscopic strategy that consists of L independent runs of the Monte Carlo EM algorithm. Let $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_L$ denote the estimates of $\boldsymbol{\theta}$ that result from the L runs, we can assess the extent of Monte Carlo variation by calculating the sample variance-covariance matrix $\mathbf{S}$ based on $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_L$. As a by-product, we can also compute

$$\overline{\boldsymbol{\theta}} = \frac{1}{L}\sum_{l=1}^{L}\widehat{\boldsymbol{\theta}}_l$$

as a more precise estimate of $\boldsymbol{\theta}$. An estimate of the asymptotic covariance matrix of $\overline{\boldsymbol{\theta}}$ that explicitly accounts for Monte Carlo variation (Kuk & Chen, 1992) is

$$V = V_1 + V_2$$

where

$$V_1 = -\{l''_M(\overline{\boldsymbol{\theta}})\}^{-1},$$

$$V_2 = \mathbf{S}/L$$

and $l''_M(\overline{\boldsymbol{\theta}})$ given by (6.19) is a Monte Carlo approximation of the observed information evaluated at $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$. If V_2 contributes negligibly to the total V, we can conclude that the Monte Carlo variation is nonsignificant.

6.2 Models for the salamander mating data

The salamander mating data reported by McCullagh and Nelder (1989, pp. 439-450) have been extensively analysed (Schall, 1991; Breslow and Clayton, 1993; Karim and Zeger, 1992). The data were recorded from experiments involving two geographically isolated populations of salamanders, Rough Butt (R) and Whiteside (W). The scientific question addressed in the study is

whether the geographically isolated species of salamanders develop barriers to successful mating. 10 R males and 10 W males were sequestered as pairs with 10 R females and 10 W females on six occasions according to the design given in Table 14.3 of McCullagh and Nelder (1989). For each pair, it was recorded whether mating occurred and there are n=360 such records altogether, 120 records from each experiment. The first experiment was conducted in the summer of 1986 while the other two experiments were conducted in the fall of the same year. The same animals were used in the first two experiments while a new set of animals was used for the third experiment. Our main objective in this analysis is to estimate the probability of a successful mating for each of the four types of cross in mating, RR (R female with R male), RW, WR, WW as well as the seasonal effect. We are also interested in knowing whether there exists heterogeneity among animals and, if so, whether it is greater for females or males and for which species. To answer this question, we model the animal effects as random effects. As the same animals were used in the first two experiments, those random effects corresponding to the same animal are correlated.

Previous analyses of the data (Schall, 1991; Breslow and Clayton, 1993; Karim and Zeger, 1992) used a generalized linear model with random effects and a logit link function. McCulloch (1994) used probit link and a Monte Carlo ECME algorithm to analyse the three experiments separately. We obtain similar results for the model proposed by McCulloch (1994, p.333), called model 0 using the Monte Carlo EM algorithm described in Section 6.1.2. For example, for experiment 1, the results of 10 runs of the algorithm are given in Table 9. We obtain $\widehat{\beta}_{rr} = 0.792(\text{SE} = 0.393)$, $\widehat{\beta}_{rw} = 0.531(0.374)$, $\widehat{\beta}_{wr} =$ -0.954(0.412), $\widehat{\beta}_{ww} = 0.698(0.385)$, $\widehat{\sigma}_f^2 = 0.592(0.363)$ and $\widehat{\sigma}_m^2 = 0.019(0.153)$ compared with the estimates 0.819, 0.538, -0.978, 0.707, 0.600 and 0.067 ob-

tained by McCulloch.

6.2.1 A probit linear model with correlated random effects

To analyse the combined data set, we use the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{f12}\mathbf{u}_{f12} + \mathbf{Z}_{m12}\mathbf{u}_{m12} + \mathbf{Z}_{f3}\mathbf{u}_{f3} + \mathbf{Z}_{m3}\mathbf{u}_{m3} + \boldsymbol{\varepsilon} \tag{6.21}$$

and $\mathbf{W} = I_{(\mathbf{Y}>0)}$. The design matrix $\mathbf{X}$ consists of the indicator variables for the four types of cross RR, RW, WR and WW as well as an indicator variable for the season (fall=1; summer=0). The vector $\boldsymbol{\beta} = (\beta_{rr}, \beta_{rw}, \beta_{wr}, \beta_{ww}, \beta_{fall})^T$ consists of the corresponding fixed effects. The random effects are

$$\mathbf{u}_{f12} \sim \mathcal{N}_{40}(\mathbf{0}, \mathbf{I}_{20} \otimes \boldsymbol{\Sigma}_{f12}), \text{where } \boldsymbol{\Sigma}_{f12} = \begin{pmatrix} \sigma^2_{f1} & \sigma_{f12} \\ \sigma_{f12} & \sigma^2_{f2} \end{pmatrix},$$

$$\mathbf{u}_{m12} \sim \mathcal{N}_{40}(\mathbf{0}, \mathbf{I}_{20} \otimes \boldsymbol{\Sigma}_{m12}), \text{where } \boldsymbol{\Sigma}_{m12} = \begin{pmatrix} \sigma^2_{m1} & \sigma_{m12} \\ \sigma_{m12} & \sigma^2_{m2} \end{pmatrix},$$

$$\mathbf{u}_{f3} \sim \mathcal{N}_{20}(\mathbf{0}, \sigma^2_{f3}\mathbf{I}_{20}),$$

$$\mathbf{u}_{m3} \sim \mathcal{N}_{20}(\mathbf{0}, \sigma^2_{m3}\mathbf{I}_{20})$$

where $\mathbf{u}_{f12}(\mathbf{u}_{m12})$ is made up of 20 2×1 random vectors corresponding to the effects of 20 female(male) animals over the 2 occasions, experiment 1 and 2 and $\mathbf{u}_{f3}(\mathbf{u}_{m3})$ represents the effects of the new set of animals used in experiment

3. The design matrices correspond to these random effects are $\mathbf{Z}_{f12}, \mathbf{Z}_{m12}, \mathbf{Z}_{f3}$ and $\mathbf{Z}_{m3}$. Finally, we assume that the error vector $\boldsymbol{\varepsilon} \sim \mathcal{N}_{360}(\mathbf{0}, \mathbf{I})$.

In fitting the model, we set the starting values to be zero for all beta parameters, 0.01 for all variance parameters and 0.001 for all covariance parameters. Using (6.6), (6.7), (6.16) and (6.17), we iterate 300 times to obtain the parameter estimates. The results of 5 runs of the algorithm are given in Table 10. It is clear that we obtain more or less the same estimates from each run. In fact, V_2 contributes only negligibly to $V = V_1 + V_2$ where $V_1 = -\{l''_M(\overline{\theta})\}^{-1}$ is stabilized at $M = 50000$.

The following conclusions are drawn. For the fixed effects, we find that the mating rate of the WR cross type is the lowest whereas the mating rates of RR and WW are the highest and of similar magnitude. The seasonal effect is in the direction of less successful mating in fall. An estimate of the contrast of primary interest $\beta_{rw} - \beta_{wr}$ is 1.397 (S.E. 0.364) which is significantly different from zero. Finally, we examine the random effects and find that the female random effects have a higher variability than the male random effects in the first two experiments whereas in the last experiment, the relation is reversed. Furthermore, the female random effects in experiment 1 and 2 are apparently not correlated while for the males, they appear to be positively correlated. These findings are in reasonable agreement with the results from previous analyses using logit link.

In assessing the goodness-of-fit, we calculate the estimated probabilities of successful mating for RR, RW, WR and WW mating types, denoted by $\pi_{rr}, \pi_{rw}, \pi_{wr}, \pi_{ww}$ in experiment 1, 2 or 3. For example,

$$\pi_{rr} = \Phi\big(\beta_{rr}/(\sigma^2_{f1} + \sigma^2_{m1} + 1)^{\frac{1}{2}}\big) \tag{6.22}$$

for experiment 1 (Zeger, Liang & Albert, 1988). Then we compare the estimated probabilities with the observed proportions in Table 12. We can see that all the estimated probabilities match quite well with the observed proportions. The sample variances of the animal specific totals for the female R, female W, male R and male W salamanders, denoted by S^2_{fr}, S^2_{fw}, S^2_{mr} and S^2_{mw} are calculated for each experiment and their expected values are approximated on the basis of 5000 samples drawn from the model at the estimated parameter values. The results, also given in Table 12, reveal the inadequacy of the model as we find that for each gender, the ordering of the expected variances for species R and W is often opposite to that of the corresponding observed variances. For example, the observed S^2_{fr} (1.733) is smaller than the observed S^2_{fw} (3.789) in experiment 2 but the expected S^2_{fr} (2.898) is larger than that of S^2_{fw} (2.138) under model 1. Therefore, we revise our model to allow the variance parameters of different species to be different.

6.2.2 A probit linear model with species specific random effects

The revised model, model 2, has 5 fixed effects and 8 random effects,

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{fr12}\mathbf{u}_{fr12} + \mathbf{Z}_{fw12}\mathbf{u}_{fw12} + \mathbf{Z}_{mr12}\mathbf{u}_{mr12} + \mathbf{Z}_{mw12}\mathbf{u}_{mw12} + \\ &\quad \mathbf{Z}_{fr3}\mathbf{u}_{fr3} + \mathbf{Z}_{fw3}\mathbf{u}_{fw3} + \mathbf{Z}_{mr3}\mathbf{u}_{mr3} + \mathbf{Z}_{mw3}\mathbf{u}_{mw3} + \boldsymbol{\epsilon}. \qquad (6.23)\end{aligned}$$

This model is an extension of model 1 by further subdividing $\mathbf{u}_{f12}$, $\mathbf{u}_{m12}$, $\mathbf{u}_{f3}$ and $\mathbf{u}_{m3}$ in (6.21) into two species specific parts. Thus

$$\mathbf{u}_{fr12} \sim \mathcal{N}_{20}(\mathbf{0}, \mathbf{I}_{10} \otimes \mathbf{\Sigma}_{fr12}),$$

$$\mathbf{u}_{fw12} \sim \mathcal{N}_{20}(\mathbf{0}, \mathbf{I}_{10} \otimes \mathbf{\Sigma}_{fw12}),$$

$$\mathbf{u}_{mr12} \sim \mathcal{N}_{20}(\mathbf{0}, \mathbf{I}_{10} \otimes \mathbf{\Sigma}_{mr12}),$$

$$\mathbf{u}_{mw12} \sim \mathcal{N}_{20}(\mathbf{0}, \mathbf{I}_{10} \otimes \mathbf{\Sigma}_{mw12}),$$

$$\mathbf{u}_{fr3} \sim \mathcal{N}_{10}(\mathbf{0}, \sigma^2_{fr3}\mathbf{I}_{10}),$$

$$\mathbf{u}_{fw3} \sim \mathcal{N}_{10}(\mathbf{0}, \sigma^2_{fw3}\mathbf{I}_{10}),$$

$$\mathbf{u}_{mr3} \sim \mathcal{N}_{10}(\mathbf{0}, \sigma^2_{mr3}\mathbf{I}_{10}),$$

$$\mathbf{u}_{mw3} \sim \mathcal{N}_{10}(\mathbf{0}, \sigma^2_{mw3}\mathbf{I}_{10})$$

where $\mathbf{u}_{fr12}(\mathbf{u}_{mr12})$ is made up of 10 2×1 random vectors corresponding to the effects of 10 female(male) animals of species R over the 2 occasions, experiment 1 and 2 and $\mathbf{u}_{fw12}(\mathbf{u}_{mw12})$ is similarily defined for species W. Again, $\mathbf{u}_{fr3}(\mathbf{u}_{mr3})$ represents the effects of the new set of species R animals used in experiment 3 and $\mathbf{u}_{fw3}(\mathbf{u}_{mw3})$ is similarily defined for species W. The design matrices correspond to these random effects are $\mathbf{Z}_{fr12}, \mathbf{Z}_{fw12}, \mathbf{Z}_{mr12}, \mathbf{Z}_{mw12}, \mathbf{Z}_{fr3}, \mathbf{Z}_{fw3}, \mathbf{Z}_{mr3}$ and $\mathbf{Z}_{mw3}$.

We use the estimates of model 1 as the starting values and iterate 400 times to obtain a new set of estimates. As our estimate of σ^2_{fw3} (0.0026) is extremely close to zero, we iterate 100 times more subject to the constraint $\sigma^2_{fw3} = 0$. We re-run the algorithm 4 times under the constraint $\sigma^2_{fw3} = 0$. The results are given in Table 11. It can be seen that the 5 runs of the Monte Carlo EM algorithm give similar results. For this example, $V_1 = -\{l''_M(\overline{\theta})\}^{-1}$ is stabilized at $M = 30000$. To assess the goodness-of-fit of model 2, we use similar procedures as for model 1 and the results are also given in Table 12. We can see that there is better agreement between the observed variances S^2_{fr}, S^2_{fw},

S^2_{mr} and S^2_{mw} and their expected values under model 2 than under model 1.

6.3 Extension

In (6.3), we allow a correlated structure for the random effects but it is assumed that the errors $\boldsymbol{\varepsilon}$ are independently and identically distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I})$. As a result, we can write down the complete data MLE in closed form which greatly simplifies the M-step of the EM algorithm. However, for longitudinal data such as the methadone clinic data, it seems more reasonable to assume that the errors for the same patient are serially correlated. To allow correlation in the error vector $\boldsymbol{\varepsilon}$, we assume that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi} \neq \mathbf{I}$. The extended model has potential applications in the analysis of longitudinal binary data and multivariate clustered binary data.

6.3.1 Extension to autocorrelated error

For longitudinal data, it is common to assume autocorrelated errors within subjects. If the time points are equally spaced, we may consider an AR(1) correlation matrix, $\Psi_i(\rho)$ given by (2.2) for subject i. Assuming independence across subjects, the correlation matrix for $\boldsymbol{\varepsilon}$ is block diagonal

$$\boldsymbol{\Psi}(\rho) = \text{diag}\big(\Psi_1(\rho), \ \ldots, \Psi_m(\rho)\big) \tag{6.24}$$

where m is the number of subjects. Under this correlation structure, the M-step of the EM algorithm becomes slightly more complicated than that in Section 6.1.1. The complete-data MLE of $\mathbf{\Sigma}_r$ is still given by (6.5) but the complete-data MLE of $\boldsymbol{\beta}$ and ρ have no closed form. For a fixed value of ρ, the MLE of $\boldsymbol{\beta}$ is the generalized least squares estimates

$$\widehat{\boldsymbol{\beta}}_c(\rho) = (\mathbf{X}^T\ \mathbf{\Psi}(\rho)^{-1}\ \mathbf{X})^{-1}\ \mathbf{X}^T\mathbf{\Psi}(\rho)^{-1}\ (\ \mathbf{Y}\ -\ \sum_{r=1}^{R}\mathbf{Z}_r\mathbf{u}_r\ ). \tag{6.25}$$

The complete-data MLE $\widehat{\rho}_c$ of ρ can be obtained by maximizing the profile log-likelihood

$$\begin{aligned} l(\rho) \ &= \ \frac{-(N-m)}{2}\log(1-\rho^2) - \\ &\frac{1}{2}\big(\mathbf{Y}\ -\ \mathbf{X}\widehat{\boldsymbol{\beta}}_c(\rho)\ -\ \sum_{r=1}^{R}\mathbf{Z}_r\mathbf{u}_r\big)^T\mathbf{\Psi}(\rho)^{-1}\big(\mathbf{Y}\ -\ \mathbf{X}\widehat{\boldsymbol{\beta}}_c(\rho)\ -\ \sum_{r=1}^{R}\mathbf{Z}_r\mathbf{u}_r\big) \end{aligned} \tag{6.26}$$

and $\widehat{\boldsymbol{\beta}}_c(\widehat{\rho}_c)$ is the complete data MLE for $\boldsymbol{\beta}$. Thus the EM algorithm works as follows. Given the current estimates $\mathbf{\Sigma}_r^{(k)}$, $\boldsymbol{\beta}^{(k)}$ and $\rho^{(k)}$, we obtain the updated estimate $\rho^{(k+1)}$ by taking conditional expectation of the above profile log-likelihood given the observed data $\mathbf{W} = I_{(\mathbf{Y}>0)}$ and maximize it with respect to ρ. The updated estimates of $\mathbf{\Sigma}_r$ and $\boldsymbol{\beta}$ are $\mathbf{\Sigma}_r^{(k+1)} = \mathrm{E}(\widehat{\mathbf{\Sigma}}_{rc}|\mathbf{W})$ and $\boldsymbol{\beta}^{(k+1)} = \mathrm{E}(\widehat{\boldsymbol{\beta}}(\rho^{(k+1)})|\mathbf{W})$ where the expectations are evaluated at the current parameter estimates.

Next, we consider the possibility of applying the AR(1) error model to the methadone data. As commented before, the probit-linear mixed model can be viewed as a threshold model that results from dichotomizing the observations from a Gaussian mixed model. As a starting point, we consider an underlying random intercept model with AR(1) errors,

$$Y_{it} = u_i + \mathbf{x}_{it}\boldsymbol{\beta} + \varepsilon_{it} \tag{6.27}$$

where u_i is i.i.d. $\mathcal{N}(0, \sigma^2)$ and

$$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix} \sim \mathcal{N}\{\mathbf{0}, \Psi_i(\rho)\}.$$

where $\Psi_i(\rho)$ is a $n_i \times n_i$ matrix given by (2.2). The observed binary data are $W_{it} = I_{(Y_{it}>0)}$. The maximum likelihood estimation of the fixed effects $\boldsymbol{\beta}$ and the variance component σ is carried out through the EM algorithm. To apply the EM algorithm, we treat the complete data as $\mathbf{Y}, u_1, \ldots, u_m$ and regard the observed data $\mathbf{W} = I_{(\mathbf{Y}>\mathbf{0})}$ as incomplete data. To obtain the complete data MLE, we could use the generalised least squares estimates defined by (6.25) but we find it more appealing to use a transformation approach. It is well known that the GLS estimates are equivalent to the OLS estimates based on the transformed data $\mathbf{Y}^* = \mathbf{P}(\mathbf{Y} - \mathbf{Zu})$, $\mathbf{X}^* = \mathbf{PX}$ and $\boldsymbol{\varepsilon}^* = \mathbf{P}\boldsymbol{\varepsilon}$ where $\mathbf{P} = \text{diag}(\mathbf{P}_1, \ldots, \mathbf{P}_m)$ is block diagonal with $\mathbf{P}_i$ defined by

$$\mathbf{P}_i(\rho) = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \ldots & 0 & 0 \\ -\rho & 1 & 0 & \ldots & 0 & 0 \\ 0 & -\rho & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & -\rho & 1 \end{pmatrix}_{n_i \times n_i} \tag{6.28}$$

and $\mathbf{P}^T\mathbf{P} = (1-\rho^2)\boldsymbol{\Psi}^{-1}$. In other words,

$$y_{i1}^* = \sqrt{1-\rho^2}\,(y_{i1} - u_i)$$

and for $t > 1$,

$$y_{it}^* = (y_{it} - u_i) - \rho\,(y_{i,t-1} - u_i)$$

and the transformation for $\mathbf{x}_{it}$ is similarily defined. The MLE of $\boldsymbol{\beta}$ under the transformed model $\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$ is the OLS estimates

$$\widehat{\boldsymbol{\beta}}_c(\rho) = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{Y}^*. \tag{6.29}$$

Substituting $\widehat{\boldsymbol{\beta}}_c(\rho)$ into (6.26), we obtain the profile log-likelihood function

$$l(\rho) = \frac{-(N-m)}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\sum_{i=1}^{m}\sum_{t=1}^{n_i}\left(y_{it}^* - \mathbf{x}_{it}^*\widehat{\boldsymbol{\beta}}_c(\rho)\right)^2 \tag{6.30}$$

which can be maximized to yield $\widehat{\rho}_c$, the complete data MLE for ρ. The complete data MLE for $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_c(\widehat{\rho}_c)$ and for σ^2, it is clear that

$$\widehat{\sigma}_c^2 = \frac{\sum_{i=1}^{m} u_i^2}{m}. \tag{6.31}$$

In the E-step, $\mathrm{E}(\mathbf{Y}_i|\mathbf{W})$, $\mathrm{E}(\mathbf{Y}_i\mathbf{Y}_i^T|\mathbf{W})$, $\mathrm{E}(u_i|\mathbf{W})$, $\mathrm{E}(u_i^2|\mathbf{W})$ and $\mathrm{E}(u_i\mathbf{Y}_i|\mathbf{W})$ are required. Equations (6.13) and (6.14) show that $\mathrm{E}(u_i|\mathbf{W})$ and $\mathrm{E}(u_i^2|\mathbf{W})$ can be expressed in terms of the conditional mean and the conditional covariance matrix of $\mathbf{Y}$ given $\mathbf{W}$. Thus $\mathrm{E}[\mathbf{Y}|\mathbf{W}]$ and $\mathrm{Cov}[\mathbf{Y}|\mathbf{W}]$ are all that we need to carry out the steps. Since the complete data MLE for $\boldsymbol{\beta}$ and ρ have no closed form, iterations are required to carry out the M-step. Thus the EM algorithm requires iterations within iterations which become very computational intensive especially for the methadone data with over two thousand observations. As a result, we do not fit this model to the methadone data.

6.3.2 Extension to multivariate clustered binary data

We consider next multivariate clustered binary data which arise, for example, in the study of multiple binary traits in animal breeding (Foulley, Gianola & Im, 1989). Let $h = 1, \ldots, H$ index the H traits. We can model each trait by a threshold model like (6.3) so that $\mathbf{W}_h = I_{(\mathbf{Y}_h > \mathbf{0})}$ and

$$\mathbf{Y}_h = \mathbf{X}_h\boldsymbol{\beta}_h + \sum_{r=1}^{R} \mathbf{Z}_{hr}\mathbf{u}_{hr} + \boldsymbol{\varepsilon}_h, \qquad h = 1, \ldots, H, \tag{6.32}$$

where $\boldsymbol{\beta}_h$ and $\mathbf{u}_{hr}$ denote the trait-specific fixed and random effects. Combining, we have

$$\left(\mathbf{Y}_1 - \sum_{r=1}^{R} \mathbf{Z}_{1r}\mathbf{u}_{1r}, \ldots, \mathbf{Y}_H - \sum_{r=1}^{R} \mathbf{Z}_{Hr}\mathbf{u}_{Hr}\right) = \left(\mathbf{X}_1\boldsymbol{\beta}_1, \ldots, \mathbf{X}_H\boldsymbol{\beta}_H\right) + \left(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_H\right). \tag{6.33}$$

The row vectors of the error matrix $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_H)$ are assumed to be independent and identically distributed as $\mathcal{N}(\mathbf{0}, \mathbf{C})$ where $\mathbf{C}$ denotes the residual covariance matrix. Note that this model is more realistic than a model which assumes that $\mathbf{Y}_1, \ldots, \mathbf{Y}_H$ are conditionally independent given the random effects $\mathbf{u}_{hr}$, $h = 1, \ldots, H$; $r = 1, \ldots, R$. Note also that the right hand side of (6.33) is a multivariate linear model. If we define the complete data as $(\mathbf{Y}_h, \mathbf{u}_{hr})$, $h = 1, \ldots, H$; $r = 1, \ldots, R$, the complete-data MLE of $\boldsymbol{\beta}_h$ and $\boldsymbol{\Sigma}_{hr}$ are just trait-specific analogues of (6.4) and (6.5). The complete-data MLE of the covariance matrix $\mathbf{C}$ is given by the sample covariance matrix

$$\widehat{\mathbf{C}} = \widehat{\boldsymbol{\varepsilon}}^T\widehat{\boldsymbol{\varepsilon}}/\, n \tag{6.34}$$

where $\widehat{\boldsymbol{\varepsilon}} = (\widehat{\boldsymbol{\varepsilon}}_1, \ldots, \widehat{\boldsymbol{\varepsilon}}_H)$ is the matrix of residuals with trait-specific residual

vector $\widehat{\boldsymbol{\varepsilon}}_h = \mathbf{Y}_h - \mathbf{X}_h\widehat{\boldsymbol{\beta}}_h - \sum_{r=1}^{R}\mathbf{Z}_{hr}\widehat{\mathbf{u}}_{hr}$. A subtle point is the following. Since only $\mathbf{W} = I_{(\mathbf{Y}>\mathbf{0})}$ is observed, the variances $c_{11}, \ldots, c_{HH}$ of the error matrix $\boldsymbol{\varepsilon}$ are not estimable from the observed data. As the scale parameter $\sqrt{c_{hh}}$ can be absorbed into the regression parameters $\boldsymbol{\beta}_h$, we can only estimate the ratios $\boldsymbol{\beta}_1/\sqrt{c_{11}}, \ldots, \boldsymbol{\beta}_H/\sqrt{c_{HH}}$. The common way to overcome this identifiability problem for probit model is to assume $c_{11} = \ldots = c_{HH} = 1$. This constraint however complicates the estimation of $\mathbf{C}$ because the constrained MLE of $\mathbf{C}$ is no longer given by (6.34). For example, if $H = 2$, it is well known that the MLE of c_{12} subject to $c_{11} = c_{22} = 1$ is the solution to a cubic equation. We are now in an interesting situation where all the parameters are estimable from the complete data $(\mathbf{Y}_h, \mathbf{u}_{hr})$, $h = 1, \ldots, H$; $r = 1, \ldots, R$ but only the ratios $\boldsymbol{\beta}_1/\sqrt{c_{11}}, \ldots, \boldsymbol{\beta}_H/\sqrt{c_{HH}}$ are estimable from the observed data $\mathbf{W} = I_{(\mathbf{Y}>\mathbf{0})}$ and we have to impose the constraint $c_{11} = \ldots = c_{HH} = 1$. To implement the EM algorithm, we think it is easier to obtain the unconstrained complete-data MLE of $\boldsymbol{\beta}_h, \boldsymbol{\Sigma}_{hr}$ and $\mathbf{C}$. The updating formula now consists of taking conditional expectation of the unconstrained MLE given the observed data and evaluating them at the current parameter estimates. We conjecture that we can obtain convergent estimates of the estimable parameters $\boldsymbol{\beta}_1/\sqrt{c_{11}}, \ldots, \boldsymbol{\beta}_H/\sqrt{c_{HH}}$ by using the above unconstrained maximization version of the EM algorithm. To our knowledge, problems where the complete-data model has more parameters than the observed data model have not been addressed before and deserve further investigation.

Appendix A

Derivatives of log-likelihood for bivariate model

Differentiating (5.8), the log-likelihood function with respect to a vector of parameters, we have

$$\frac{\partial l}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{m}\sum_{t=1}^{n_i}\Big[\, Y_{ith}Y_{itb}\; P_{it}(1,1)^{-1}\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_j} \;+\; Y_{ith}(1-Y_{itb})\; P_{it}(1,0)^{-1}\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_j} \;+$$
$$(1-Y_{ith})Y_{itb}\; P_{it}(0,1)^{-1}\tfrac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_j} \;+\; (1-Y_{ith})(1-Y_{itb})\; P_{it}(0,0)^{-1}\tfrac{\partial P_{it}(0,0)}{\partial \boldsymbol{\theta}_j}\Big] \quad \text{(A.1)}$$

for $j = h, b$ or r where

$$\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_h} = \frac{\mathbf{X}_{ith}S^2_{ith}}{2}\Big[1-\Delta_{ith}\delta_{it}^{-\frac{1}{2}}\Big], \quad \text{(A.2)}$$

$$\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_b} = \frac{\mathbf{X}_{itb}S^2_{itb}}{2}\Big[1-\Delta_{itb}\delta_{it}^{-\frac{1}{2}}\Big], \quad \text{(A.3)}$$

$$\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_r} = \frac{\mathbf{X}_{itr}\gamma_{it}}{2}\Big\{\Big[-\Delta_{itr}\delta_{it}^{-\frac{1}{2}} - (1-\delta_{it}^{\frac{1}{2}})(\gamma_{it}-1)^{-1}\Big]\,(\gamma_{it}-1)^{-1}\Big\}, \quad \text{(A.4)}$$

$$\Delta_{ith} = \gamma_{it}[P_{it}(1,\cdot) - P_{it}(\cdot,1)] - [P_{it}(1,\cdot) + P_{it}(\cdot,1)] + 1,$$

$$\Delta_{itb} = -\gamma_{it}[P_{it}(1,\cdot) - P_{it}(\cdot,1)] - [P_{it}(1,\cdot) + P_{it}(\cdot,1)] + 1,$$

$$\Delta_{itr} = \gamma_{it}[P_{it}(1,\cdot) - P_{it}(\cdot,1)]^2 - [P_{it}(1,\cdot)^2 + P_{it}(\cdot,1)^2] + [P_{it}(1,\cdot) + P_{it}(\cdot,1)],$$

$$\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_h} = \mathbf{X}_{ith} S_{ith}^2 - \frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_h};$$

$$\frac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_b} = \mathbf{X}_{itb} S_{itb}^2 - \frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_b};$$

$$\frac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_h} = -\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_h}; \frac{\partial P_{it}(0,0)}{\partial \boldsymbol{\theta}_h} = -\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_h};$$

$$\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_b} = -\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_b}; \frac{\partial P_{it}(0,0)}{\partial \boldsymbol{\theta}_b} = -\frac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_b};$$

$$\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_r} = \frac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_r} = -\frac{\partial P_{it}(0,0)}{\partial \boldsymbol{\theta}_r} = -\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_r},$$

$S_{ith}^2 = P_{it}(1,\cdot)[1 - P_{it}(1,\cdot)] = e^{\eta_{ith}}/(1+e^{\eta_{ith}})^2$, $S_{itb}^2 = P_{it}(\cdot,1)[1 - P_{it}(\cdot,1)]$ $= e^{\eta_{itb}}/(1+e^{\eta_{itb}})^2$, $\gamma_{it} = e^{\eta_{itr}}$ and $\mathbf{X}_{itj}$ is a $q_j \times 1$ vector, $j = h, b$ or r. Again, the second order derivatives of the log-likelihood function are

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_l^T} = \sum_{i=1}^{m}\sum_{t=1}^{n_i} \Big\{ & Y_{ith}Y_{itb}\left[P_{it}(1,1)^{-1}\frac{\partial^2 P_{it}(1,1)}{\partial \boldsymbol{\theta}_j \boldsymbol{\theta}_l^T} - P_{it}(1,1)^{-2}\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_j}\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_l^T}\right] + \\
& Y_{ith}(1-Y_{itb})\left[P_{it}(1,0)^{-1}\frac{\partial^2 P_{it}(1,0)}{\partial \boldsymbol{\theta}_j \boldsymbol{\theta}_l^T} - P_{it}(1,0)^{-2}\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_j}\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_l^T}\right] + \\
& (1-Y_{ith})Y_{itb}\left[P_{it}(0,1)^{-1}\frac{\partial^2 P_{it}(0,1)}{\partial \boldsymbol{\theta}_j \boldsymbol{\theta}_l^T} - P_{it}(0,1)^{-2}\frac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_j}\frac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_l^T}\right] + \\
& (1-Y_{ith})(1-Y_{itb})\left[P_{it}(0,0)^{-1}\frac{\partial^2 P_{it}(0,0)}{\partial \boldsymbol{\theta}_j \boldsymbol{\theta}_l^T} - P_{it}(0,0)^{-2}\frac{\partial P_{it}(0,0)}{\partial \boldsymbol{\theta}_j}\frac{\partial P_{it}(0,0)}{\partial \boldsymbol{\theta}_l^T}\right]\Big\}
\end{aligned} \tag{A.5}$$

for $j, l = h, b$ or r where

$$\frac{\partial^2 P_{it}(1,1)}{\partial \boldsymbol{\theta}_h \partial \boldsymbol{\theta}_h^T} = \frac{\mathbf{X}_{ith}\mathbf{X}_{ith}^T}{2}\left\{ A_{ith} + (S_{ith}^2)^2(\gamma_{it}-1)\Delta_{ith}^2\,\delta_{it}^{-\frac{3}{2}} - \left[(S_{ith}^2)^2(\gamma_{it}-1)+\Delta_{ith}A_{ith}\right]\delta_{it}^{-\frac{1}{2}}\right\}, \tag{A.6}$$

$$\frac{\partial^2 P_{it}(1,1)}{\partial \boldsymbol{\theta}_b \partial \boldsymbol{\theta}_b^T} = \frac{\mathbf{X}_{itb}\mathbf{X}_{itb}^T}{2}\left\{ A_{itb} + (S_{itb}^2)^2(\gamma_{it}-1)\Delta_{itb}^2\,\delta_{it}^{-\frac{3}{2}} - \left[(S_{itb}^2)^2(\gamma_{it}-1)+\Delta_{itb}A_{itb}\right]\delta_{it}^{-\frac{1}{2}}\right\}, \tag{A.7}$$

$$\frac{\partial^2 P_{it}(1,1)}{\partial \boldsymbol{\theta}_h \partial \boldsymbol{\theta}_b^T} = \frac{\mathbf{X}_{ith}\mathbf{X}_{itb}^T}{2}\left\{ S_{ith}^2 S_{itb}^2\left[\,(\gamma_{it}-1)\Delta_{ith}\Delta_{itb}\delta_{it}^{-\frac{3}{2}} + \,(\gamma_{it}+1)\,\delta_{it}^{-\frac{1}{2}}\right]\right\}, \tag{A.8}$$

$$\frac{\partial^2 P_{it}(1,1)}{\partial \boldsymbol{\theta}_h \partial \boldsymbol{\theta}_r^T} = \frac{\mathbf{X}_{ith}\mathbf{X}_{itr}^T}{2}\left\{ S_{ith}^2\gamma_{it}\left[\Delta_{ith}\Delta_{itr}\delta_{it}^{-\frac{3}{2}} \; - \; [P_{it}(1,\cdot) - P_{it}(\cdot,1)]\,\delta_{it}^{-\frac{1}{2}}\right]\right\}, \tag{A.9}$$

$$\frac{\partial^2 P_{it}(1,1)}{\partial \boldsymbol{\theta}_b \partial \boldsymbol{\theta}_r^T} = \frac{\mathbf{X}_{itb}\mathbf{X}_{itr}^T}{2}\left\{ S_{itb}^2\gamma_{it}\left[\Delta_{itb}\Delta_{itr}\delta_{it}^{-\frac{3}{2}} \; + \; [P_{it}(1,\cdot) - P_{it}(\cdot,1)]\,\delta_{it}^{-\frac{1}{2}}\right]\right\}, \tag{A.10}$$

$$\begin{aligned}\frac{\partial^2 P_{it}(1,1)}{\partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_r^T} \;=\; & \frac{\mathbf{X}_{itr}\mathbf{X}_{itr}^T}{2}\left\{\gamma_{it}\gamma_{it}\left\{\,\left[-[P_{it}(1,\cdot) - P_{it}(\cdot,1)]^2\delta_{it}^{-\frac{1}{2}} + \Delta_{itr}^2\delta_{it}^{-\frac{3}{2}}\right] \; + \right.\right. \\ & \left.\left.\left[2(\gamma_{it}-1)^{-1} - \gamma_{it}^{-1}\right]\left[\Delta_{itr}\delta_{it}^{-\frac{1}{2}} + (1-\delta_{it}^{\frac{1}{2}})(\gamma_{it}-1)^{-1}\right]\right\}(\gamma_{it}-1)^{-1}\right\},\end{aligned} \tag{A.11}$$

$$\frac{\partial^2 P_{it}(1,0)}{\partial \boldsymbol{\theta}_h \partial \boldsymbol{\theta}_h^T} = \mathbf{X}_{ith}\mathbf{X}_{ith}^T A_{ith} - \frac{\partial^2 P_{it}(1,1)}{\partial \boldsymbol{\theta}_h \partial \boldsymbol{\theta}_h^T};$$

$$\frac{\partial^2 P_{it}(0,1)}{\partial \boldsymbol{\theta}_b \partial \boldsymbol{\theta}_b^T} = \mathbf{X}_{itb}\mathbf{X}_{itb}^T A_{itb} - \frac{\partial^2 P_{it}(1,1)}{\partial \boldsymbol{\theta}_b \partial \boldsymbol{\theta}_b^T};$$

$$\frac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_h \partial \boldsymbol{\theta}_h^T} = -\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_h \partial \boldsymbol{\theta}_h^T}; \frac{\partial P_{it}(0,0)}{\partial \boldsymbol{\theta}_h \partial \boldsymbol{\theta}_h^T} = -\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_h \partial \boldsymbol{\theta}_h^T};$$

$$\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_b \partial \boldsymbol{\theta}_b^T} = -\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_b \partial \boldsymbol{\theta}_b^T}; \frac{\partial P_{it}(0,0)}{\partial \boldsymbol{\theta}_b \partial \boldsymbol{\theta}_b^T} = -\frac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_b \partial \boldsymbol{\theta}_b^T};$$

$$\frac{\partial P_{it}(1,0)}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_l^T} = \frac{\partial P_{it}(0,1)}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_l^T} = -\frac{\partial P_{it}(0,0)}{\partial \boldsymbol{\theta}_j \boldsymbol{\theta}_l^T} = -\frac{\partial P_{it}(1,1)}{\partial \boldsymbol{\theta}_j \boldsymbol{\theta}_l^T},$$

for $j \neq l$ or $j = l = r$, $A_{ith} = P_{it}(1,\cdot) - 3P_{it}(1,\cdot)^2 + 2P_{it}(1,\cdot)^3$ and $A_{itb} = P_{it}(\cdot,1) - 3P_{it}(\cdot,1)^2 + 2P_{it}(\cdot,1)^3$.

Table

Table 1 *Parameter estimates and standard errors (in italic) for various AR(1) models.*

model	β_o	β_d	β_t	β_{p1}
marginal	-0.2195	-0.0108	-0.3435	n.a.
($\alpha = 0.4305$)	*0.3769*	*0.00569*	*0.0736*	n.a
conditional	-0.8423	-0.00884	-0.4049	2.3960
	0.2189	*0.00282*	*0.0628*	*0.1196*
random β_o(GQ$_4$)	-0.9280	-0.00685	-0.4277	1.5516
	0.4286	*0.00761*	*0.0704*	*0.1420*
random β_o (ML)	-0.5859	-0.0132	-0.4100	1.4038
	0.3410	*0.00490*	*0.0682*	*0.1334*
random β_o, β_d	-0.1057	-0.0180	-0.3869	n.a.
	0.4058	*0.00610*	*0.0692*	n.a.
random β_o, β_d	-0.4777	-0.0149	-0.4130	1.4018
	0.3339	*0.00497*	*0.0682*	*0.1334*

Table 2 *Parameter estimates and standard errors (in italic) for various AR(2) models.*

model	β_o	β_d	β_t	β_{p1}	β_{p2}
conditional	-0.8978 *0.2215*	-0.00870 *0.00286*	-0.4477 *0.0643*	1.9802 *0.1306*	1.1013 *0.1393*
random β_o(GQ$_4$)	-0.8909 *0.3808*	-0.00730 *0.00620*	-0.3787 *0.0699*	1.4884 *0.1453*	0.4943 *0.1536*
random β_o (ML)	-0.6153 *0.3282*	-0.0128 *0.00471*	-0.4305 *0.0684*	1.3311 *0.1382*	0.4535 *0.1478*
random β_o, β_d	-0.5001 *0.3203*	-0.0146 *0.00478*	-0.4338 *0.0683*	1.3299 *0.1382*	0.4513 *0.1478*

Table 3 *Parameter estimates for separate fitting to selected patients.*

i.d.	β_o	β_d	β_t	β_{p1}
103	15.2154	-0.2394	0.6600	1.4362
119	6.4219	-0.1919	1.8743	1.0175
120	4.1748	-0.0831	-0.6182	0.9896
123	13.8571	-0.1036	-2.2498	1.1565
132	-27.9921	0.4709	0.4176	-2.2771
134	0.6367	0.0292	-1.4076	0.8603
164	-5.5810	0.1491	-0.4587	0.1173
172	-0.8535	0.0290	-0.7472	1.7880
173	-6.9913	0.1437	-0.8278	0.9546
509	0.8920	0.0250	-0.9770	-1.7253
512	5.5039	-0.1501	1.0544	1.7088
513	5.7359	-0.1264	-0.3930	0.4757
517	7.5551	-0.0258	-2.2858	0.0491
522	0.1196	0.0280	-1.2460	-0.0220
523	5.6147	-0.0606	-1.0315	0.0266
532	0.6642	-0.0392	0.4130	0.5868
538	-2.9978	-0.4569	7.6776	-2.7581
540	44.7475	-0.8159	-5.8932	2.5568
567	2.1490	-0.0139	-1.6775	0.8289
573	-0.8423	-0.1507	3.9988	0.4805
574	-26.4234	0.1123	7.1960	-0.3037
575	4.9208	-0.0122	-1.4421	-0.1700
587	9.7951	-1.3778	23.9124	-0.3560
589	152.7	-0.8755	-37.4085	-0.2922

Table 4 *Standardized Score statistics and P-value (in italic) for testing homogeneity of each coefficient.*

	β_o	β_d	β_t	β_{p1}
Standardized Score	6.2391	4.9716	5.4184	2.3915
P-value	*0.0000*	*0.0000*	*0.0000*	*0.0084*

Table 5 *Parameter estimates, standard errors (in italic) and AIC for mixture models with different number of groups.*

model	β_o	β_d	β_t	β_{p1}	π	L	AIC
$G = 2$ (group 1)	-1.1732 *0.5165*	-0.0153 *0.00692*	-0.4317 *0.0686*	1.5610 *0.1371*	0.6689 *0.0571*	-977.37	1968.75
(group 2)	-0.1321 *0.3100*	0.00097 *0.00421*	same	same	0.3311 *0.0571*		
$G = 3$ (group 1)	-5.2306 *2.0002*	0.0249 *0.0242*	-0.4262 *0.0694*	1.4313 *0.1397*	0.3039 *0.0785*	-967.45	1954.89
(group 2)	0.1314 *0.3922*	-0.0245 *0.00603*	same	same	0.4603 *0.0846*		
(group 3)	0.1305 *0.4008*	0.00047 *0.00528*	same	same	0.2358 *0.0500*		
$G = 4$ (group 1)	-5.2110 *1.8800*	0.0258 *0.0232*	-0.4233 *0.0705*	1.3886 *0.1426*	0.3104 *0.0780*	-966.37	1958.75
(group 2)	0.3120 *0.4670*	-0.0275 *0.00737*	same	same	0.4449 *0.0842*		
(group 3)	0.5743 *0.5381*	-0.00707 *0.00778*	same	same	0.1853 *0.0665*		
(group 4)	-1.3950 *1.0416*	0.0270 *0.0169*	same	same	0.0594 *0.0559*		
$G = 5$						-966.29	1964.58

Table 6 *Estimated group memberships $\widehat{W}_{ik}$ for the two-group and three-group mixture AR(1) models.*

			2-group		3-group		
Pat.No.	n_i	$y_{i\cdot}$	1	2	1	2	3
103	26	7	0.02	0.98	0.00	0.32	0.68
104	26	1	1.00	0.00	0.48	0.51	0.00
105	10	3	0.49	0.51	0.00	0.76	0.24
106	10	0	0.98	0.02	0.78	0.21	0.01
109	26	2	0.98	0.02	0.14	0.86	0.00
110	15	8	0.01	0.99	0.00	0.25	0.75
111	6	1	0.71	0.29	0.13	0.69	0.18
113	4	4	0.02	0.98	0.00	0.11	0.89
118	26	5	0.61	0.39	0.00	0.93	0.07
119	26	18	0.00	1.00	0.00	0.00	1.00
120	26	3	0.96	0.04	0.01	0.99	0.01
121	26	0	1.00	0.00	0.92	0.08	0.00
122	26	0	1.00	0.00	0.97	0.03	0.00
123	14	10	0.00	1.00	0.00	0.00	1.00
124	26	0	1.00	0.00	0.69	0.31	0.00
125	4	0	0.93	0.07	0.46	0.50	0.03
126	21	1	1.00	0.01	0.34	0.66	0.00
128	5	0	0.93	0.07	0.67	0.30	0.03
129	11	0	0.99	0.01	0.70	0.29	0.00
131	26	2	0.98	0.02	0.12	0.87	0.00
132	15	7	0.00	1.00	0.00	0.06	0.94
133	9	4	0.17	0.83	0.00	0.55	0.45
134	22	10	0.00	1.00	0.00	0.02	0.98
135	11	4	0.25	0.75	0.00	0.64	0.36
137	8	2	0.63	0.37	0.03	0.76	0.20
138	26	0	1.00	0.00	0.69	0.31	0.00
143	26	0	1.00	0.00	0.90	0.10	0.00
144	26	6	0.13	0.87	0.00	0.55	0.45
145	26	0	1.00	0.00	0.85	0.15	0.00
146	6	3	0.28	0.72	0.00	0.56	0.44
149	26	1	1.00	0.00	0.59	0.41	0.00
150	9	4	0.23	0.77	0.00	0.64	0.36
153	12	1	0.95	0.05	0.22	0.75	0.03
156	26	1	1.00	0.00	0.45	0.55	0.00
159	26	0	1.00	0.00	0.97	0.03	0.00
160	6	2	0.72	0.28	0.00	0.81	0.19
161	26	9	0.00	1.00	0.00	0.03	0.97
162	21	1	0.99	0.01	0.41	0.58	0.00
163	26	4	0.89	0.11	0.00	0.97	0.03
164	26	10	0.00	1.00	0.00	0.26	0.74

			2-group		3-group		
Pat.No.	n_i	$y_{i\cdot}$	1	2	1	2	3
165	25	0	1.00	0.00	0.68	0.32	0.00
166	12	1	0.93	0.07	0.19	0.79	0.03
168	26	2	0.99	0.01	0.11	0.89	0.00
169	26	0	1.00	0.00	0.87	0.13	0.00
170	15	8	0.00	1.00	0.00	0.06	0.94
171	26	5	0.42	0.58	0.00	0.92	0.08
172	26	14	0.00	1.00	0.00	0.00	1.00
173	26	15	0.00	1.00	0.00	0.00	1.00
175	14	7	0.00	1.00	0.00	0.08	0.92
501	24	8	0.00	1.00	0.00	0.00	1.00
502	25	5	0.21	0.79	0.00	0.75	0.25
503	13	0	0.98	0.02	0.95	0.05	0.00
504	26	0	1.00	0.00	0.83	0.17	0.00
505	26	0	1.00	0.00	0.79	0.21	0.00
506	26	0	1.00	0.00	0.95	0.05	0.00
507	26	1	1.00	0.00	0.46	0.54	0.00
508	26	2	0.98	0.02	0.16	0.84	0.00
509	26	8	0.04	0.96	0.00	0.66	0.34
510	26	0	1.00	0.00	0.75	0.25	0.00
511	26	0	1.00	0.00	0.92	0.08	0.00
512	10	3	0.34	0.66	0.00	0.69	0.31
513	26	11	0.00	1.00	0.00	0.21	0.79
514	12	1	0.95	0.05	0.45	0.54	0.01
515	18	0	1.00	0.00	0.46	0.54	0.00
516	26	2	0.99	0.01	0.04	0.96	0.00
517	26	11	0.00	1.00	0.00	0.00	1.00
518	10	0	0.98	0.02	0.69	0.31	0.00
519	26	0	1.00	0.00	0.95	0.05	0.00
520	26	1	1.00	0.00	0.39	0.61	0.00
521	26	0	1.00	0.00	0.79	0.21	0.00
522	12	6	0.01	0.99	0.00	0.11	0.89
523	26	5	0.30	0.70	0.00	0.85	0.15
524	26	1	1.00	0.00	0.37	0.63	0.00
525	26	0	1.00	0.00	0.84	0.16	0.00
526	26	7	0.13	0.87	0.00	0.81	0.19
527	26	3	0.94	0.06	0.04	0.95	0.01
528	26	1	1.00	0.00	0.25	0.75	0.00
529	26	1	1.00	0.00	0.38	0.62	0.00
530	26	3	0.96	0.04	0.01	0.98	0.01
531	7	3	0.03	0.97	0.09	0.03	0.88
532	13	5	0.09	0.91	0.00	0.58	0.42
533	26	5	0.77	0.23	0.00	0.94	0.06
534	26	5	0.62	0.38	0.00	0.95	0.05
535	5	0	0.94	0.06	0.52	0.45	0.02
536	26	1	1.00	0.00	0.41	0.59	0.00
537	6	0	0.94	0.06	0.71	0.27	0.02
538	17	4	0.58	0.42	0.00	0.91	0.09
539	26	2	0.99	0.01	0.18	0.82	0.00
540	15	8	0.01	0.99	0.00	0.21	0.79
541	26	1	1.00	0.00	0.40	0.60	0.00

			2-group		3-group		
Pat.No.	n_i	$y_{i\cdot}$	1	2	1	2	3
542	26	0	1.00	0.00	0.83	0.17	0.00
543	17	11	0.00	1.00	0.00	0.00	1.00
544	26	0	1.00	0.00	0.87	0.13	0.00
545	26	12	0.00	1.00	0.00	0.11	0.89
546	26	0	1.00	0.00	0.97	0.03	0.00
547	26	1	1.00	0.00	0.45	0.55	0.00
548	26	0	1.00	0.00	0.86	0.14	0.00
549	26	1	1.00	0.00	0.38	0.62	0.00
550	26	0	1.00	0.00	0.80	0.20	0.00
551	26	1	1.00	0.00	0.10	0.89	0.00
552	26	2	0.97	0.03	0.11	0.88	0.01
553	8	0	0.98	0.02	0.48	0.52	0.01
554	26	1	1.00	0.00	0.30	0.70	0.00
555	26	9	0.04	0.96	0.00	0.68	0.32
556	10	5	0.04	0.96	0.00	0.31	0.69
557	26	2	0.98	0.02	0.15	0.85	0.00
558	26	0	1.00	0.00	0.71	0.29	0.00
559	4	0	0.94	0.06	0.36	0.61	0.03
560	26	5	0.32	0.68	0.00	0.86	0.14
561	26	1	1.00	0.00	0.44	0.56	0.00
562	10	3	0.58	0.42	0.00	0.81	0.19
563	26	0	1.00	0.00	0.49	0.51	0.00
564	18	3	0.82	0.18	0.01	0.81	0.18
565	26	6	0.27	0.73	0.00	0.86	0.14
566	13	2	0.88	0.12	0.02	0.94	0.05
567	26	3	0.93	0.07	0.09	0.90	0.01
568	18	0	1.00	0.00	0.74	0.26	0.00
570	26	3	0.96	0.04	0.01	0.98	0.01
571	26	0	1.00	0.00	0.58	0.42	0.00
572	26	2	0.99	0.01	0.18	0.81	0.00
573	26	13	0.00	1.00	0.00	0.00	1.00
574	26	3	0.96	0.04	0.00	0.99	0.01
575	26	17	0.00	1.00	0.00	0.00	1.00
576	26	21	0.00	1.00	0.00	0.00	1.00
577	16	10	0.00	1.00	0.00	0.00	1.00
579	26	0	1.00	0.00	0.94	0.06	0.00
580	26	1	1.00	0.00	0.35	0.64	0.00
581	26	0	1.00	0.00	0.79	0.21	0.00
582	24	0	1.00	0.00	0.57	0.43	0.00
583	26	2	0.98	0.02	0.20	0.80	0.00
584	17	9	0.00	1.00	0.00	0.00	1.00
585	11	0	0.98	0.02	0.91	0.09	0.00
586	26	4	0.83	0.17	0.00	0.97	0.03
587	24	4	0.59	0.41	0.00	0.93	0.07
588	26	0	1.00	0.00	0.48	0.52	0.00
589	26	12	0.00	1.00	0.00	0.04	0.96

Table 7 *Parameter estimates and standard errors (in italic) for various bivariate binary models.*

Model	β_o	β_d	β_t	β_{ph}	β_{pb}	β_{ph*b}	β_c	L (AIC)
1. H	-0.9253	-0.00714	-0.3909	2.3590	-0.4679	0.4258		-1880.92
	0.2205	*0.00292*	*0.0629*	*0.1269*	*0.2402*	*0.3799*		(3797.83)
B	-3.9962	0.0307	-0.2838	-0.0895	3.0038	-0.8412		
	0.2931	*0.00339*	*0.0731*	*0.2265*	*0.1459*	*0.3921*		
R	1.8803	-0.0253	0.0951	-1.0064	-1.0623	2.2225		
	0.7127	*0.00915*	*0.1938*	*0.5536*	*0.5365*	*0.9272*		
2. H	-0.9313	-0.00731	-0.3920	2.4092	-0.3036			-1881.64
	0.2205	*0.00292*	*0.0628*	*0.1196*	*0.1842*			(3795.28)
B	-4.0007	0.0307	-0.2821	-0.0903	2.9958	-0.8429		
	0.2929	*0.00339*	*0.0730*	*0.2264*	*0.1458*	*0.3942*		
R	2.0250	-0.0252		-0.9421	-1.0103	2.1513		
	0.6407	*0.00915*		*0.5396*	*0.5314*	*0.9170*		
3. H	-0.9324	-0.00730	-0.3916	2.4091	-0.3020			-1881.69
	0.2203	*0.00291*	*0.0628*	*0.1196*	*0.1841*			(3791.39)
B	-4.0060	0.0308	-0.2812	-0.0905	2.9955	-0.8298		
	0.2924	*0.00339*	*0.0730*	*0.2264*	*0.1458*	*0.3913*		
R	1.0163	-0.0248					1.0156	
	0.7597	*0.00897*					*0.4108*	

Table 8 *Parameter estimates and standard errors (in italic) for various bivariate binary mixture models.*

Model	β_o	β_d	β_t	β_{ph}	β_{pb}	β_{ph*b}	β_c	π	L (AIC)
4. H 1	-1.0401	-0.00859	-0.5888	1.6278	-0.3779			0.5495	-1763.03
	0.5209	*0.00735*	*0.1101*	*0.3390*	*0.8723*			*0.0512*	(3578.06)
2	-0.3195	-0.00662	-0.2969	2.0606	-0.9579			0.4505	
	0.2957	*0.00376*	*0.0886*	*0.1623*	*0.1997*			*0.0512*	
B 1	-5.6318	0.0361	-0.3293	0.8113	2.2951	-1.2327			
	0.8361	*0.00915*	*0.1860*	*0.5430*	*0.4566*	*1.3260*			
2	-3.0432	0.0332	-0.1896	-1.2857	1.8513	0.4198			
	0.3751	*0.00437*	*0.0874*	*0.2638*	*0.1807*	*0.4287*			
R	0.8315	-0.0245					0.5185		
	0.7551	*0.00966*					*0.4102*		
5. H 1	-1.1156	-0.00759	-0.5981	1.9576	-0.5124			0.5558	-1764.93
	0.3562	*0.00362*	*0.1119*	*0.1389*	*0.8405*			*0.0516*	(3569.86)
2	-0.2599	same	-0.2767	same	-0.9564			0.4442	
	0.2893		*0.0903*		*0.1949*			*0.0516*	
B 1	-5.3743	0.0335	-0.3289	0.5221	1.9345				
	0.5197	*0.00386*	*0.1810*	*0.5105*	*0.1577*				
2	-3.0618	same	-0.2060	-1.1180	same				
	0.3496		*0.0873*	*0.2066*					
R	0.7685	-0.0236					0.5648		
	0.7454	*0.00957*					*0.4035*		
6. H 1	-1.5301	-0.00750	-0.3977	1.9587	-0.3021			0.5547	-1767.24
	0.3102	*0.00365*	*0.0652*	*0.1358*	*0.7742*			*0.0507*	(3570.47)
2	-0.0081	same	same	same	-0.9704			0.4453	
	0.2790				*0.1954*			*0.0507*	
B 1	-5.5876	0.0334	-0.2248	0.6741	1.9504				
	0.3852	*0.00390*	*0.0785*	*0.4894*	*0.1563*				
2	-3.0220	same	same	-1.1412	same				
	0.3414			*0.2080*					
R	0.6868	-0.0224					0.5639		
	0.7441	*0.00955*					*0.4024*		

Table 9 *Parameter estimates of the probit-linear model for experiment 1 salamander data.*

Model 0	Fixed effects				Var. of rand. effects	
	β_{rr}	β_{rw}	β_{wr}	β_{ww}	σ^2_{f1}	σ^2_{m1}
	0.7921	0.5278	-0.9481	0.6917	0.5816	0.0195
	0.7882	0.5287	-0.9517	0.6899	0.5783	0.0190
	0.7943	0.5345	-0.9546	0.7060	0.6013	0.0194
	0.7963	0.5302	-0.9466	0.6967	0.5917	0.0190
	0.7933	0.5258	-0.9502	0.6964	0.5923	0.0194
	0.7881	0.5320	-0.9569	0.7020	0.5951	0.0193
	0.7911	0.5318	-0.9578	0.7025	0.5880	0.0195
	0.7883	0.5326	-0.9505	0.7003	0.5868	0.0190
	0.7901	0.5305	-0.9635	0.7024	0.6071	0.0194
	0.7935	0.5335	-0.9588	0.6964	0.5995	0.0194
Average	0.7915	0.5308	-0.9539	0.6984	0.5922	0.0193
V_2	*8.10E-7*	*7.26E-7*	*2.85E-6*	*2.61E-6*	*7.99E-6*	*4.30E-9*
$V = V_1 + V_2$	*0.1546*	*0.1396*	*0.1695*	*0.1483*	*0.1321*	*0.0235*
$SE = V^{\frac{1}{2}}$	*0.3932*	*0.3736*	*0.4117*	*0.3852*	*0.3634*	*0.1532*

Table 10 *Parameter estimates of the probit-linear model with correlated random effects for the salamander data.*

Model 1	Fixed effects							
	β_{rr}	β_{rw}	β_{wr}	β_{ww}	β_{fall}			
	0.8739	0.4426	-0.9484	0.8373	-0.3532			
	0.8787	0.4464	-0.9545	0.8384	-0.3566			
	0.8809	0.4480	-0.9524	0.8432	-0.3614			
	0.8735	0.4421	-0.9536	0.8362	-0.3523			
	0.8761	0.4445	-0.9508	0.8414	-0.3549			
Average	0.8766	0.4447	-0.9519	0.8393	-0.3557			
V_2	*2.01E-6*	*1.26E-6*	*1.17E-6*	*1.71E-6*	*2.55E-6*			
$V = V_1 + V_2$	*0.1106*	*0.1000*	*0.1086*	*0.1048*	*0.0883*			
$SE = V^{\frac{1}{2}}$	*0.3326*	*0.3162*	*0.3296*	*0.3237*	*0.2972*			
	Variance and covariance of random effects							
Species	σ^2_{f1}	σ^2_{f2}	σ_{f12}	σ^2_{m1}	σ^2_{m2}	σ_{m12}	σ^2_{f3}	σ^2_{m3}
	0.6301	0.7033	-0.0419	0.2683	0.4776	0.3556	0.1455	0.6758
	0.6338	0.7137	-0.0396	0.2709	0.4825	0.3591	0.1441	0.6750
	0.6338	0.7211	-0.0446	0.2674	0.4762	0.3544	0.1488	0.6617
	0.6311	0.6993	-0.0453	0.2704	0.4818	0.3585	0.1460	0.6797
	0.6444	0.7133	-0.0511	0.2704	0.4814	0.3584	0.1477	0.6794
Average	0.6347	0.7102	-0.0445	0.2695	0.4799	0.3572	0.1464	0.6743
V_2	*6.53E-6*	*1.53E-5*	*3.73E-6*	*4.68E-7*	*1.59E-6*	*8.48E-7*	*6.87E-7*	*1.08E-5*
$V = V_1 + V_2$	*0.1718*	*0.2067*	*0.0735*	*0.0246*	*0.0824*	*0.0439*	*0.0393*	*0.1530*
$SE = V^{\frac{1}{2}}$	*0.4145*	*0.4547*	*0.2712*	*0.1569*	*0.2870*	*0.2095*	*0.1982*	*0.3912*

Table 11 *Parameter estimates of the probit-linear model with species specific random effects for the salamander data.*

Model 2	Fixed effects							
	β_{rr}	β_{rw}	β_{wr}	β_{ww}	β_{fall}			
	0.8938	0.4180	-1.0060	0.8641	-0.3872			
	0.8917	0.4131	-1.0083	0.8708	-0.3830			
	0.8906	0.4228	-1.0108	0.8666	-0.3879			
	0.8907	0.4183	-1.0018	0.8700	-0.3890			
	0.8935	0.4185	-1.0078	0.8713	-0.3891			
Average	0.8921	0.4181	-1.0070	0.8685	-0.3873			
V_2	*4.68E-7*	*2.32E-6*	*2.27E-6*	*1.93E-6*	*1.23E-6*			
$V = V_1 + V_2$	*0.1081*	*0.0832*	*0.1478*	*0.1065*	*0.0950*			
$SE = V^{\frac{1}{2}}$	*0.3288*	*0.2885*	*0.3844*	*0.3264*	*0.3082*			
	Variance and covariance of random effects							
Species	σ^2_{f1}	σ^2_{f2}	σ_{f12}	σ^2_{m1}	σ^2_{m2}	σ_{m12}	σ^2_{f3}	σ^2_{m3}
R	0.4648	0.2026	-0.1990	0.3734	0.9081	0.5807	0.5133	0.7262
	0.4531	0.1976	-0.1930	0.3831	0.9287	0.5949	0.5268	0.7231
	0.4536	0.2043	-0.1995	0.3781	0.9179	0.5875	0.5180	0.7371
	0.4505	0.1956	-0.1927	0.3780	0.9184	0.5876	0.5162	0.7221
	0.4494	0.1985	-0.1930	0.3734	0.9069	0.5804	0.5123	0.7333
Average	0.4543	0.1997	-0.1954	0.3772	0.9160	0.5862	0.5173	0.7284
V_2	*7.54E-6*	*2.61E-6*	*2.45E-6*	*3.24E-6*	*1.57E-5*	*7.14E-6*	*6.68E-6*	*8.65E-6*
$V = V_1 + V_2$	*0.2179*	*0.0375*	*0.0563*	*0.0794*	*0.4332*	*0.1802*	*0.2261*	*0.3668*
$SE = V^{\frac{1}{2}}$	*0.4668*	*0.1937*	*0.2372*	*0.2818*	*0.6582*	*0.4245*	*0.4755*	*0.6057*
W	0.8440	1.8853	0.2278	0.1202	0.1402	0.1290	0.0000	0.8166
	0.8683	1.8804	0.2196	0.1176	0.1369	0.1260	0.0000	0.8767
	0.8657	1.8355	0.2124	0.1217	0.1419	0.1307	0.0000	0.8450
	0.8535	1.7916	0.2168	0.1190	0.1387	0.1277	0.0000	0.8230
	0.8753	1.8331	0.2282	0.1193	0.1390	0.1280	0.0000	0.8450
Average	0.8613	1.8452	0.2209	0.1196	0.1394	0.1283	0.0000	0.8413
V_2	*3.12E-5*	*2.98E-4*	*9.61E-6*	*4.79E-7*	*7.11E-7*	*5.84E-7*	*0.00E0*	*1.12E-4*
$V = V_1 + V_2$	*0.4448*	*2.4047*	*0.3523*	*0.0106*	*0.0274*	*0.0201*	*0.0000*	*0.4572*
$SE = V^{\frac{1}{2}}$	*0.6669*	*1.5507*	*0.5935*	*0.1031*	*0.1654*	*0.1419*	*0.0000*	*0.6762*

Table 12 *Observed and expected values for various statistics.*

	Observed proportion			Model 0	Model 1			Model 2		
				Expected proportion						
	Exp 1	Exp 2	Exp 3	Exp 1	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3
π_{rr}	0.733	0.600	0.667	0.734	0.737	0.638	0.650	0.745	0.636	0.632
π_{rw}	0.667	0.467	0.533	0.662	0.626	0.524	0.526	0.631	0.511	0.508
π_{wr}	0.233	0.233	0.167	0.226	0.245	0.188	0.166	0.250	0.236	0.144
π_{ww}	0.700	0.667	0.633	0.709	0.728	0.628	0.640	0.731	0.610	0.639
	Observed variance			Expected variance						
	Exp 1	Exp 2	Exp 3	Exp 1	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3
S^2_{fr}	2.844	1.733	2.711	2.671	2.574	2.898	1.714	2.283	1.944	2.406
S^2_{fw}	2.844	3.789	0.711	2.263	2.159	2.138	1.285	2.399	3.282	1.001
S^2_{mr}	2.100	2.944	2.278	1.102	1.523	1.788	2.148	1.652	2.398	2.021
S^2_{mw}	1.878	1.378	3.833	1.261	1.801	2.412	3.137	1.497	1.725	3.319

Bibliography

Abramowitz, M. and Stegun, I. (1972) *Handbook of Mathematical Functions.* New York: Dover.

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics & Computing*, **6**, 251-262

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

Azzalini, A. (1994) Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, **81**, 767-775.

Ball, J.C. & Ross, A. (1991) *The Effectiveness of Methadone Maintenance Treatment: Patients, Programs, Services and Outcome.* New York: Springer-Verlag.

Bock, D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, **46**, 443-459.

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

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9-25.

Commenges, D., Letenneur, L., Jacqmin, H., Moreau, T. and Dartigues, J. (1994) Test of homogeneity of binary data with explanatory variables. *Biometrics*, **50**, 613-620.

Crouch, E.A.C. and Spiegelman, D. (1990) The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)dt$: application to logistic-normal models. *J. Am. Statist. Ass.*, **85**, 464-469.

Drum, M.L. and McCullagh, P. (1993) REML estimation with exact covariance in the logistic mixed model. *Biometrics*, **49**, 677-689.

Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions.* New York: Wiley.

Foulley, J.L., Gianola, D. and Im, S. (1989) Genetic evaluation for discrete polygenic traits in animal breeding, In *Advances in Statistical Methods for Genetic Improvement of Livestock*, ed. Gianola, D. and Hammond, K., Springer-Verlag, pp. 361-411.

Lefkopoulou, M., Moore, D. and Ryan, L. (1989) The analysis of multiple correlated binary outcomes: application to rodent teratology experiments. *J. Am. Statist. Assoc.*, **84**, 810-815.

Laird, N. (1978) Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *J. Am. Statist. Assoc.*, **73**, 805-811.

Liang, K.Y. and Zeger, S.L. (1989) A class of logistic regression models for multivariate binary time series. *J. Am. Statist. Ass.*, **84**, 447-451.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992) Multivariate regression analyses for categorical data (with discussion). *J. R. Statist. Soc. B*, **54**, 3-40.

Lindsay, B.G. (1983) The Geometry of Mixture Likelihoods: A General Theory. *Ann. Statist.*, **11**, No.1, 86-94.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical analysis with missing data.* New York: Wiley.

Liu, C. and Rubin, D.B. (1994) The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, **81**, 633-648.

Louis, T.A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. B*, **44**, 226-233.

Karim, M.R. and Zeger, S.L. (1992) Generalized linear models with random effects; salamander data revisited. *Biometrics*, **48**, 631-644.

Kuk, A.Y.C. (1995) Asymptotically unbiased estimation in generalized linear models with random effects. *J. R. Statist. Soc. B*, **57**, 395-407.

Kuk, A.Y.C. and Chen, C.H. (1992) A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**, 531-541.

Marsaglia, G. (1964) Generating a variable from the tail of the normal distribution. *Technometrics*, **6**, 101-102.

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

McCulloch, C.E. (1994) Maximum likelihood variance components estimation for binary data. *J. Am. Statist. Ass.*, **89**, 330-335.

McGilchrist, C.A. (1994) Estimation in generalized mixed models. *J. R. Statist. Soc. B*, **56**, 61-69.

Meng, X.L. and Rubin, D.B. (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Ass.*, **86**, 899-909.

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

Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719-727.

Smith, A.F.M. and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *J. R. Statist. Soc. B*, **55**, 3-23.

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

Waclawiw, M.A. and Liang, K.Y. (1993) Prediction of random effects in the generalized linear model. *J. Am. Statist. Ass.*, **88**, 171-178.

Wei, G.C.G. and Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Statist. Ass.*, **85**, 699-704.

Zeger, S.L. and Karim, M.R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *J. Am. Statist. Ass.*, **86**, 79-86.

Zeger, S.L. and Liang, K.Y. (1991) Feedback models for discrete and continuous time series. *Statistica Sinica*, **1**, 51-64.

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