Models for Correlated Binary Responses: Applications for the Waterloo Smoking Prevention Projects Data

by

Andreas Istvan Sashegyi

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 1998

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

This thesis discusses several modelling approaches for the analysis of correlated binary data, as motivated by the results of a longitudinal school-based smoking prevention trial. In this study, WSPP3 - the third in a series called the Waterloo Smoking Prevention Projects, longitudinal observations were collected on a cohort of students, randomized to various study conditions in clusters defined by schools.

An extension to the logistic-normal empirical Bayes random effects model is proposed, termed quasi empirical Bayes (QEB), which allows for the estimation of fixed effect model parameters, adjusting simultaneously for the effects of extraneous school-to-school variability as well as intra-individual correlation. A method of generating data with a composite correlation structure similar to that of the Smoking Prevention Projects data is developed; this is subsequently implemented in a simulation study to examine the properties of parameter estimates from the QEB model.

We then move into a discussion of the relationship between marginal or population-averaged models and cluster-specific models, and show that a straightforward modification of cluster-specific random effects models can be used to approximate a marginal correlation structure. With this approach one could for example determine easily whether or not the intra-school correlation in the WSPP3 data depends on school size. Our discussion focusses on the logistic-normal model in particular, with estimation in this case proceeding by maximum likelihood. Power considerations and the effects of model misspecification are addressed.

Finally, we consider a general approach for testing the fit of models for correlated data. The emphasis here is on assessing how well a particular model captures the covariance structure of the data, assuming the mean is correctly specified. Some analytic results are given, as well as ones based on simulation.

Acknowledgements

I would like to extend heartfelt thanks to my supervisor, Dr. Stephen Brown, for his constant support and patient guidance throughout the entire course of my studies. I am also very grateful for having had the benefit of co-supervision from Dr. Patrick Farrell, who on numerous occasions went to the effort of traversing half this continent to meet with me in person. It was a privilege to have been able to work under Drs. Brown and Farrell. Further, I am deeply indebted to Drs. R. J. O'Hara-Hines and J. F. Lawless, whose many helpful suggestions greatly improved the quality of the final draft of this thesis.

I also thank Cheryl Madill, Robyn Landers and Neil Patterson for their support in data management and computing issues. Their constant willingness to be of assistance was exemplary.

During the past number of years I have received the care and support of many individuals, all of whom have contributed to making my time at Waterloo a very positive experience. I owe a great deal:

To my parents, who celebrated my successes with me and stood behind me during setbacks; who prayed for me.

To my brother and his family, who I look up to, who heartened me constantly and whose doors were always open to me.

To my mentor and friend, Robert Liddy, whose steadfast affirmation and unshakeable belief in my abilities taught me a powerful lesson in faith.

To Robert Zuccherato, with whom I share so many common experiences and whose support and sound advice benefitted me so often.

To Robert Reinhart, whose recent friendship greatly impacted on and enriched my final

months at Waterloo.

I regret that space prevents me from mentioning in like manner numerous other friends and relatives; I thank them all.

Further, I would like to express my gratitude to Drs. Brown and Farrell, the Department of Statistics and Actuarial Science, the University of Waterloo and the Natural Sciences and Engineering Research Council of Canada for the generous financial support which I have received over the years.

I end by thanking, above all, God, without Whose help this thesis could not have been written.

Andreas Sashegyi
University of Waterloo
April 1998

To my brother Hubert

Omnia in philosophia, omnes in philosopho continentur

Victor Hugo

Notre-Dame de Paris

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Correlated Data

Correlated data arise whenever the mechanism which produces a set of observations causes these observations to be dependent on one another. This is a fairly common phenomenon which occurs in various settings. For instance data from a time series, a split-plot experimental design or a longitudinal study all display some form of correlation.

This work focusses on methods for the analysis of clustered data. A set of data is said to be clustered if it can be logically divided into a number of groups such that it is reasonable to assume that observations are correlated within, but not between, these groups or clusters. Note that we assume here that only a single such grouping of the data is possible; if there is more than one sense in which the data are clustered, a more general definition must be adopted. (This problem is addressed in Chapter 4).

Clustering in data, typically leading to overdispersion (or, in the case of binary responses, extra-binomial variation), can occur for a variety of reasons. Sometimes one or more outlying observations are responsible for introducing extraneous variance into the data, which is eliminated once these points are removed, downweighted, or otherwise ex-

plained in the analysis. A set of data may also appear overdispersed if certain important cluster-level covariates have been omitted in an analysis. Often these covariates constitute unknown factors, which, if they were available and accounted for, would remove the apparent cluster effect. This situation is quite common and may be resolved with the use of a random effects model (see section 3.3.3). Random effects may be thought of as a means of adjusting an analysis for all the 'missing' covariates which would be necessary to allow a reasonably accurate modelling of the response. Finally, correlation within a cluster is frequently due to some common factor shared by and influencing all respondents within that cluster. Examples of such factors include litter effects in epidemiological studies involving laboratory animals, where the clusters are litters of rat pups, say, and correlation of responses within a litter is due to the fact that each pup in the litter comes from the same mother; environmental factors might play a similar role in studies where groups of subjects are randomized and studied as units; finally, in longitudinal studies repeated observations are made over time on each of many individuals, in which case individuals themselves yield clusters of observations, which are correlated since they come from the same person. Conditional models (section 3.3.2) or marginal models (section 3.3.1) are best equipped to handle intra-cluster correlation of this type.

The last example in the preceding paragraph illustrates the generality of the notion of a 'cluster'. There are at least two subjectively different senses in which this term can be applied. The first refers to groups of individuals which by virtue of some factor can be considered a more or less cohesive unit. The term cross-sectional clusters will often be used to describe such groups. In contrast, studies carried out over a period of time give rise to longitudinal clusters; in this case a cluster corresponds to all the observations taken on a single individual over time. It seems apparent that observations within such a longitudinal cluster should be correlated, whereas the clusters (individuals) themselves might be assumed to be independent.

1.2 Overview and Scope

This thesis focusses primarily on the development of new methods for analyzing correlated binary data, and also on an approach for assessing the goodness-of-fit of models for such data. The Waterloo Smoking Prevention Project, Phase 3 (WSPP3), a school-based smoking prevention trial, will be used to motivate the necessity for these methods and also to demonstrate their application. Chapter 2 gives a description of this study. Broadly speaking, interest is focussed on examining factors associated with the onset of smoking in adolescent children, and on related issues, such as appropriately reflecting the correlation structure in the data. Some specific questions are outlined in section 2.2.

Numerous methods have been proposed for the analysis of clustered data, both for continuous and discrete responses. An overview of these is given in Chapter 3. We begin with a general discussion of the interpretation of regression coefficients in population-averaged versus cluster-specific models, and describe three different model types for correlated data. There follows a discussion of several methods for dealing with overdispersed binary data, including generalized estimating equations (GEEs) and random effects models.

Existing methods, however, do not adequately address the problem of modelling data in which there is both longitudinal correlation as well as cross-sectional clustering. This is witnessed in the WSPP3, wherein observations are collected over time on each of a cohort of students attending, at any given time point, a variety of schools. Clearly one should anticipate that the smoking status of a particular student at one time point is related to his or her response at the previous time point(s), hence repeated observations on the same student will be correlated. At the same time the fact that all students in a given school are subject to the same environment will likely cause additional correlation among observations from the same school. Chapter 4 describes an approach for dealing with such a composite dependence structure, using a combination of empirical Bayes and GEE methodology. The proposed model is described in section 4.2 and simulation results to assess the properties

of estimators are given in section 4.5.

In Chapter 5 we consider cross-sectionally clustered binary data more carefully, and examine extensions of standard random effects models which can be used to approximate a marginal correlation structure. This more general class of random effects models can incorporate covariate information thought to be associated with the correlation structure in a straightforward and natural way. A parameter γ is specified and estimated, which determines a relationship between covariate(s) and the covariance struture of the data in much the same way as the fixed effects parameters establish a connection between covariates and the mean response. Furthermore, in contrast to GEE, this modelling framework admits likelihood-based analyses. The connection between cluster-specific and population-averaged models is of direct relevance here, and is discussed in section 5.3. The following section deals with issues related to the choice of particular model formulations. In section 5.5 we discuss the results of simulations carried out to address questions concerning the power of tests of hypotheses, specifically tests of $H_0 : \gamma = 0$ vs $H_A : \gamma > 0$, and the effects of model misspecification in this context.

We address the problem of assessing the goodness-of-fit of models for correlated data in Chapter 6. Interest here is focussed on how well the estimated covariance structure from a given model fit reflects the empirical covariance structure of the data. To this end, we propose standardized measures for assessing model fit based on the *a priori* assumption that the mean response function of the model has been correctly specified. The performance of this approach is illustrated in section 6.3, using some of the simulated data from Chapter 5. Section 6.4 provides an example illustrating the usefulness of the goodness-of-fit measures described in analyzing real data.

Chapter 7 is devoted to some illustrations of the methods presented in this thesis, using the WSPP3 data. Concluding remarks are given in the final chapter.

Chapter 2

The Waterloo Smoking Prevention Projects Data

2.1 Overview

The methods discussed in this thesis will be illustrated using primarily data from the Waterloo Smoking Prevention Project 3 (WSPP3), the third in a series of randomized, controlled smoking prevention trials, designed to develop, evaluate and disseminate an effective school-based social influences smoking prevention program (see Best et al. (1995) and Brown and Cameron (1997)).

This study consisted of an elementary as well as a highschool component, enrolling a total of approximately 6000 students. Initially 100 elementary schools from seven school boards were ranked high, medium or low, according to their smoking-associated risk. This ranking was based on smoking rates of older students within each school. Stratifying on both risk score and school board, the schools were randomized to one of five study conditions. Four of these were treatment conditions, corresponding to the 4 combinations of the type of provider who administered the intervention curriculum (nurse or teacher)

and the type of training the provider received (workshop or mediated training through printed material). The fifth was a control condition; students enrolled in schools in this category received only their school's existing health education program. Starting in grade 6, students were exposed each year until grade 8 to the smoking prevention curriculum, after which they moved on into secondary schools. Two variables of particular interest that were collected during the delivery of the curriculum were scores of content and style. These were designed to measure respectively the proportion of the planned curriculum which was actually covered in class, and the style in which it was delivered. A baseline measure of smoking status was taken prior to any intervention at the beginning of grade 6, and subsequently smoking status was measured on the same students at the end of grades 7 and 8.

As part of the highschool component of this study, the students of the WSPP3 elementary cohort were followed to the end of grade 12, and their smoking status measured on an annual basis in grades 9 through 12. In addition, 30 schools, each of which enrolled 30 or more students from the original cohort, were matched in pairs according to location (urban versus rural), size, and the proportion of cohort students from elementary school intervention conditions. The schools in each pair were then randomized to either an intervention or a control condition. The highschool intervention program covered the period to the end of grade 10 for the cohort, and consisted of a school mobilization effort to involve students in activities supportive of non-smoking. Systematic attempts were initiated by a selected staff member in each school to maximize such student participation in promoting the smoke-free cause (Brown and Cameron (1997)).

2.2 Some Questions of Interest

There are a number of questions one might wish to investigate using the WSPP3 data. Each of these address one or more of three fundamental queries, which we list below:

1. Is the intervention effective?

2. Can assertions be made regarding the onset of smoking?

3. How does the intervention mediate or relate to smoking onset?

The second of these questions is concerned more with the incidence of smoking or the rate at which students begin smoking, than with smoking prevalence, i.e. the proportion of smokers at a given time. The third question investigates the difference in the effect of the intervention among subgroups of the study population. Various modelling approaches are suitable for addressing these issues, a number of which are described in the subsequent chapters.

In the WSPP3 study a categorical variable was used to classify a student's smoking status. The values 1 through 5 were assigned respectively to the categories 'never smoked', 'smoked once', 'quit', 'experimental smoker' and 'regular smoker'. Quitters were defined to be individuals who had smoked more than once but who considered themselves as having 'quit'; experimental smokers were those currently smoking less frequently than once a week and regular smokers those smoking once a week or more. Many questions of interest can be addressed by dichotomizing this 5-point scale, the cut-point being determined by the nature of the inquiry. For example classifying students with responses 1, 2 or 3 as non-smokers and those with responses 4 or 5 as smokers is appropriate when focussing on smoking status at a given point in time. In contrast, if one is interested in making assertions about the time to smoking onset (time until the first smoking experience), one would group students on the basis of response values of 1 versus 2 or greater.

Initially one might consider fitting cross-sectional models, modelling the responses of students separately at each grade. Marginal response models (see section 3.3.1), which regress the (unconditional) outcome variable on a series of predictor variables, are appropriate for this purpose. These models are useful if one is interested in examining, at

a particular point in time, the population-averaged effect of predictor variables on the response (see also section 3.2).

Alternatively, one might wish to carry out a longitudinal survival-type analysis and model, as indicated above, time to first smoking, using a survival model of the type described in section 3.5.3. In this case data from several grades are concatenated as follows: begin by considering non-smokers only at the initial time point (grade 6), and at each subsequent time point (grade) retain only those observations in the analysis corresponding to non-smokers in the previous grade. Since smoking prevention is one of the primary goals of the Waterloo Smoking Prevention Projects, such an analysis is relevant and meaningful.

Conditional or more specifically transition models as discussed in sections 3.3.2 and 3.5.4 are useful if one is interested in describing transitions from one state to another, such as non-smoking to smoking and vice versa. Such models have an important place since smoking status at time point $t-1$ is highly predictive of the response at time t; it is therefore sensible to model the response at time t conditional on the response at previous time point(s). One can distinguish a variety of specific models which fall into this context. A Markov model, for instance, would be appropriate if a student's response at a given time depended on his/her behaviour at the previous time point, but were independent of earlier responses. (Interestingly, there is evidence in the WSPP3 data to suggest that in fact smoking status is strongly influenced by the observations made at the previous two time points). When dealing with a small number of repeated responses over time, it is actually feasible to estimate the full joint distribution of the repeated observations, by using a fully conditional model which expresses the joint distribution in terms of a product of conditionals.

Evaluating the effectiveness of a smoking prevention program such as that implemented in WSPP3 is not a straightforward task, since there are many perspectives from which one can view the problem. One must first be clear about the specific questions one wishes to investigate about the 'treatment', using this term in a generic sense, before one can

make any conclusions about a treatment effect. For example, an analysis which examines not only smoking *rates*, but conditional on these also investigates the *amount* smoked, might discover that the rates are largely the same in both treated and untreated groups, but that individuals in the treated group smoke less on average. In this case one could still argue in favour of a positive treatment effect. One might also find that a certain treatment combination helps to reduce smoking rates in students, but only for a subgroup of the original cohort. For example, the intervention program may have a positive effect on students in high-risk groups (defined by family or school environment), but not on other students. Differential treatment effects might appear in different clusters, or in subgroups defined by additional covariates under consideration.

As indicated in Chapter 1 the study design of WSPP3 gives rise to repeated observations on individual students, which are also correlated cross-sectionally, with clusters defined by schools. One goal is to facilitate marginal longitudinal analyses, incorporating random school effects to capture the overdispersion due to the disparity in smoking rates between schools. The impact of school environment is largely considered a nuisance here, but the random school effects in such models enable us to obtain asymptotically correct standard errors for regression parameters (see Chapter 3) and to identify outliers, in the sense of schools with extreme smoking rates (low or high), having adjusted for relevant covariates.

Finally, in some cases we might wish to examine behaviour patterns among students more carefully in order to better understand the smoking phenomenon as it pertains to classes or schools as a whole. It is well known that students sampled within a given school tend to exhibit similar smoking behaviour, as compared to a sample of students from different schools. One might however be interested in assessing in addition whether or not intra-school correlation, or equivalently school-to-school variability, is a function of school size. A plausible and likely explanation as to why this might be the case is a theory attributing the correlation between observations in the same school to the very strong behavioural similarity among students in small peer groups within a given school.

The larger the school, the more such subgroups one would have and hence the weaker the overall intra-school correlation (see section 7.1.2). On the level of the individual it would be of interest to assess whether anything can be said about the similarity between the smoking behaviours of two students, based on gender or the environment each has been a part of. Questions such as these can be addressed by models which allow covariate information to be incorporated into the specification of the correlation structure. Chapter 5 will present a useful framework for this purpose, and examples are discussed in Chapter 7.

Chapter 3

Models and Methods of Analysis for Correlated Binary Data

3.1 Introduction

Most standard methods of analysis applied to a set of data rely on the statistical assumption that the observations comprising the data are independent. However as the previous discussion has pointed out, common study designs often violate this assumption, and hence methods are required which sytematically take account of the special correlation structure which might be present in a given set of data. Standard methods applied to clustered data will typically lead to incorrect inferences; this point is made clear when considering the estimates of variability obtained for the regression coefficients from a model fit. When a model is fit to clustered data under the assumption that all observations are independent, variance estimates are usually underestimated, leading to inflated test statistics and hence possible conclusions which are not warranted (see for example McCullagh and Nelder (1989) and Liang and Zeger (1986); section 3.4.1 also provides a discussion of this point with reference to extra-binomial variation). To illustrate, consider a logistic model for the

grade 8 smoking rates observed from the WSPP3 cohort. Describing the probability of smoking as a function of a content score (indicating how much of the grade 8 prevention curriculum was actually covered), an indicator for type of training received by the curriculum provider ($1 \equiv$ trained through a workshop, $0 \equiv$ trained via mediated material), and an indicator of previous (grade 7) smoking status, yields an estimated regression coefficient of -0.186 for the training indicator, with associated standard error of 0.113. Hence there is weak evidence to suggest that workshop-trained providers effect lower smoking rates than those trained through mediated material, having adjusted for the other terms in the model. However, fitting the same logistic model but also accounting for the correlation expected between students in the same school, using GEE (see section 3.5), results in a point estimate of -0.149 with standard error 0.170. Therefore the training variable is no longer a significant term in the model, having adjusted for the intra-cluster correlation.

Methods for correlated data incorporate the dependence structure of the data into the analysis for either or both of the following reasons: the dependence structure may be of scientific interest in itself, and accounting for it in an appropriate manner will help ensure that valid standard errors and conclusions will be obtained. When the response variable is approximately Gaussian, a wide range of methods for the analysis of longitudinal or otherwise clustered data is available. Laird and Ware (1982) for example give a very lucid description of random-effects models for longitudinal data, and Ware (1985) presents an overview of linear models for normally distributed longitudinal data. Our focus however is on discrete outcomes, in particular binary responses. Hence the proposed methodology will combine the theory underlying generalized linear models (GLMs) and quasi-likelihood (Nelder and Wedderburn (1972), Wedderburn (1974), McCullagh (1983), McCullagh and Nelder (1989)) with some form of modelling of the correlation structure in the data.

3.2 Interpretation of Regression Coefficients

Neuhaus et al. (1991) point out that most of the approaches which have been developed for the analysis of clustered binary data can be categorized into two groups, exemplifying either a cluster-specific (CS) or a population-averaged (PA) approach. Suppose our data consist of observations from K clusters, with n_k responses observed in the kth cluster. Let Y_{ij} denote the response of the jth individual in cluster i, $j = 1, ..., n_i$ and $i = 1, ..., K$, and associate with each Y_{ij} a single covariate x_{ij}. (Extending the following discussion to vector valued x_{ij} is straightforward). As described in Neuhaus et al. (1991), in the cluster-specific approach, the response Y_{ij} is modelled as a function of the covariate x_{ij} as well as parameters α_i specific to each cluster. Hence the regression parameter β_{cs} from such a model measures a cluster-specific effect of the covariates x_{ij} on Y_{ij}. Consider for example a mixed-effects logistic model for the binary response Y_{ij}, having the form

$$\text{logit}P(Y_{ij} = 1|\alpha_i, x_{ij}) = \alpha_i + \beta_{cs}x_{ij}. \tag{3.2.1}$$

In this model β_{cs} measures the change in the log odds of the probability of response with a unit change in the covariate x_{ij}, but only conditional on the cluster-specific effect α_i. In contrast, under the population-averaged approach, the marginal expectation of Y_{ij} (averaged over the population) is modelled as a function of the covariate x_{ij} only. Thus a model of this type might take the form

$$\text{logit}P(Y_{ij} = 1|x_{ij}) = \alpha + \beta_{pa}x_{ij}, \tag{3.2.2}$$

where β_{pa} now measures the change in the log odds of $P(Y_{ij} = 1)$ with the covariate x_{ij}, unconditionally for all subjects. That is, β_{pa} measures a population-averaged effect.

In the mixed-effects model (3.2.1) the clustered nature of the data is reflected in the model through the cluster-specific parameters α_i. Model (3.2.2) on the other hand accounts

for the intra-cluster correlation through the specification of some working covariance structure for $Y_i = (Y_{i1}, ..., Y_{in_i})^T$; hence this is an example of a GEE model (see Liang and Zeger (1986) and section 3.5).

Although cluster-specific and population-averaged parameters have different interpretations, a simple approximate relationship exists between them. In equation (3.2.1) it is assumed that the terms α_i vary between clusters according to a distribution with density $g(\alpha)$. Hence taking the expectation of $P(Y_{ij} = 1|\alpha_i, x_{ij})$ with respect to α yields a unique marginal model as in equation (3.2.2), since

$$\begin{aligned} \mathrm{E}(P(Y_{ij} = 1|\alpha_i, x_{ij})) &= \int (1 + e^{-\alpha - \beta_{cs} x_{ij}})^{-1} g(\alpha) d\alpha \\ &= P(Y_{ij} = 1|x_{ij}). \end{aligned} \tag{3.2.3}$$

Omitting subscripts and referring to equation (3.2.2), it is easily shown that

$$\beta_{pa} = \log\{\frac{P(Y = 1|x + 1)/P(Y = 0|x + 1)}{P(Y = 1|x)/P(Y = 0|x)}\} \tag{3.2.4}$$

which does not depend on x. From (3.2.3) and (3.2.4) we see that β_{cs} and β_{pa} are related through the equation

$$\beta_{pa}(x) = \log\{\frac{\mathrm{E}[(1 + e^{-\alpha - \beta_{cs}(x+1)})^{-1}]\mathrm{E}[(1 + e^{\alpha + \beta_{cs} x})^{-1}]}{\mathrm{E}[(1 + e^{\alpha + \beta_{cs}(x+1)})^{-1}]\mathrm{E}[(1 + e^{-\alpha - \beta_{cs} x})^{-1}]}\} \tag{3.2.5}$$

which does depend on x. It turns out however that a linear Taylor series approximation of (3.2.5), taken about $\beta_{cs} = 0$, is independent of x: let $f(\beta_{cs})$ denote the right-hand side of equation (3.2.5). Then for small β_{cs},

$$\beta_{pa}(x) = f(\beta_{cs}) \simeq f(0) + f'(0)\beta_{cs}, \tag{3.2.6}$$

where the dash represents differentiation with respect to β_{cs}. Now $f(0) = 0$, and some

straightforward algebra shows that

$$f'(\beta_{cs})|_{\beta_{cs}=0} = \frac{(x+1)\mathrm{E}(pq)\mathrm{E}(q) - x\mathrm{E}(pq)\mathrm{E}(p) + (x+1)\mathrm{E}(pq)\mathrm{E}(p) - x\mathrm{E}(pq)\mathrm{E}(q)}{\mathrm{E}(p)\mathrm{E}(q)}$$
$$= \frac{\mathrm{E}(pq)\mathrm{E}(q) + \mathrm{E}(pq)\mathrm{E}(p)}{\mathrm{E}(p)\mathrm{E}(q)}, \qquad (3.2.7)$$

where $p = e^{\alpha}/(1+e^{\alpha})$ and $q = 1-p$, which is independent of x. Simplifying (3.2.7) and substituting the result into (3.2.6) yields the approximation

$$\beta_{pa}(x) \simeq \beta_{cs}\{1 - \frac{\mathrm{Var}(p)}{\mathrm{E}(p)\mathrm{E}(q)}\}. \qquad (3.2.8)$$

Since $p > 0$ and $\mathrm{Var}(p) \leq \mathrm{E}(p)[1-\mathrm{E}(p)] = \mathrm{E}(p)\mathrm{E}(q)$, it follows that $0 \leq \mathrm{Var}(p)/\mathrm{E}(p)\mathrm{E}(q) \leq 1$; (see also section 3.4.1). This suggests that at least for small parameter values, β_{pa} is smaller in magnitude than β_{cs}. Neuhaus et al. (1989) show analytically that this is true in general. Furthermore, if $\mathrm{Var}(p) = 0$, or equivalently if α is not a random effect, but fixed, then $\beta_{pa} = \beta_{cs}$ as one would expect.

Whether one should use a population-averaged or a cluster-specific model in any given circumstance will largely depend on the application. This question is essentially equivalent to that of deciding between a marginal and a conditional model, and in many cases considerable debate persists as to which analysis is more appropriate (see for example Lindsey and Lambert (1997) and the discussion in the next section). Adding to the difficulty is the fact that some analyses cannot be classified as purely population-averaged or cluster-specific, but rather incorporate features of both approaches; this is especially true when more than one form of clustering is present in the data. Examples include for instance the survival models discussed in section 3.5.3 or the transition models in section 3.5.4, and certainly the composite modelling approach introduced in Chapter 4. Generally, however, if a model for the marginal probability of response is of primary interest, and the clustering inherent in the data can be regarded as a nuisance, it is appropriate to use a population-averaged

approach. On the other hand, in longitudinal studies for example, in which multiple observations are made over a period of time on each of several subjects, so that subjects form clusters, it may also be of interest to obtain estimates of changes within individuals over time. These can only be obtained from a cluster-specific model. At the same time however, one should note that effective use of cluster-specific models is limited by the amount of information available per subject (cluster). If only a few observations are made on each subject it may not be possible to accurately estimate the cluster-specific parameters α_i; see Zeger et al. (1988).

3.3 Model Types

Virtually all modelling approaches can be categorized as belonging to at least one of three classes: the marginal, conditional or random effects models. The distinctive features of each are summarized below.

3.3.1 Marginal Models

Much of the literature discussing marginal models does so with reference to longitudinal data; we shall adopt the same convention. Diggle et al. (1994) provides a very suitable general reference. The application of generalized estimating equations in this context is well documented; see for example Liang and Zeger (1986), Zeger et al. (1988) and Liang et al. (1992). A critical review of marginal models is provided in Lindsey and Lambert (1997), and an interesting cautionary note is given in Pepe and Anderson (1994).

The primary focus of marginal models is on a regression function describing how the population-averaged response, not that of any one individual, depends on a set of covariates. One is mainly concerned with the parameters for the marginal expectation of the response. Accordingly, the dependence structure among the repeated observations on a gi-

ven subject is of secondary interest. As indicated in Zeger and Liang (1992), this approach can be viewed as the natural analogue for correlated data, of standard generalized linear models (GLM's) or quasi-likelihood methods for independent data; the regression coefficients from this approach also have the same interpretation. The specification of a marginal model follows that of any GLM in that a regression relationship is postulated between the marginal mean of the response and the explanatory covariates, via some link function; we further assume that the marginal variance is a function of the marginal mean. To complete the specification, some form of correlation is assumed between the repeated observations on the same individual, and this is modelled by one or more additional parameters. These however have no bearing on the interpretation of the regression coefficients.

Section 2.2 mentioned fitting cross-sectional models to the WSPP3 data, modelling the responses of students separately at each grade. This would give us information about factors affecting the smoking rates of students at one particular point in time. A logical extension would be to fit a marginal model to the combined data from several grades, giving us a single global and generally more useful summary of the effect of such factors, using all available data. Again, although there are advantages to choosing a sensible form for the correlation structure which we assume for the response vector of a given individual, this is typically not of primary interest.

3.3.2 Conditional Models

Carrying on in the context of longitudinal data, conditional models, in contrast to marginal models, focus simultaneously on the parameters for the mean specification of the response and the intra-individual correlations. One may refer again to Diggle et al. (1994), or for example to papers by Prentice (1988) and Zeger and Liang (1992). Generally speaking, any model which is not marginal can be considered to be conditional in some sense. However here we discuss in particular fully conditional, or transition models, which have special

relevance for longitudinal data in which there is a natural ordering of the observations in time. Instead of describing the marginal expectation of an individual's response at a given time, a model is postulated for the conditional expectation, given previous responses as well as relevant covariates. In this way, the pairwise correlations between repeated observations on the same individual are implicitly determined. This is perhaps the most natural way to proceed in the analysis of repeated measurements data, given the temporal ordering, especially if one is interested in assessing not a population-averaged effect, but that for a specific individual. This conditional approach takes into account in an explicit manner the individual histories of subjects, something the marginal model largely neglects. There is of course a trade-off in that the population-averaged effect discussed above is obscured in a conditional analysis. It is fair to say that useful information can be drawn from both approaches. Thus for example, in assessing the impact of a smoking prevention curriculum on students' smoking behaviour, one will likely be interested in the effect of the curriculum on average smoking rates in schools, as well as in looking at the impact on particular students, with various different individual profiles. It seems unlikely that a single analysis would be sufficient to achieve both ends to satisfaction.

3.3.3 Random Effects Models

Random effects models are based on the explicit identification of individual and population characteristics; i.e. it is assumed that the response of each individual is influenced by effects specific to that individual as well as by effects common to the whole population. In this sense the term 'mixed effects model' is perhaps more appropriate. This type of model is another example of the cluster-specific approach. It is precisely illustrated by equation (3.2.1); note that when a cluster refers to a group of observations on the same individual such a model is normally termed subject-specific.

Random or mixed effects models assume that responses from the same individual (or the

same cluster) are independent, given that individual's (or cluster's) specific coefficient(s). Using the notation of section 3.2, we assume for example that Y_{ij} and $Y_{ij'}$ are independent given α_i. Unconditionally, with respect to the whole population, this induces correlation between observations from the same individual or cluster; in the case of binary data this correlation will also depend on the mean formulation. Generally, such models are not likely to be very suitable for describing longitudinal data, in which the correlation between two repeated observations normally depends on the time between them. They are more effective in modelling clustered data arising by design or through some external factor (i.e. the heterogeneity due to cross-sectional clustering, as described in section 1.1). In the WSPP3 data, the primary sampling units (schools) are subsampled to obtain responses of individual students. This design naturally lends itself to postulating random school effects to capture the impact of school environment on an individual's response.

Another interpretation and justification for the use of random effects which is often applicable is in terms of latent variables. We assume that all responses within a cluster are affected by the same unobserved realization of some underlying random variable. Thus for example, we postulate some sort of effect on students' smoking behaviour due to school environment, without directly observing this effect. If the data consist of relatively few clusters, one can adjust for the presence of such a latent variable by including in a model a separate indicator variable for each cluster; this then replaces the random effects model by a fixed effects model. However as the number of clusters increases it becomes more convenient to assume that the unobserved effects are derived from some distribution.

Laird and Ware (1982) provide a good discussion of random effects models in the linear framework, and Stiratelli, Laird and Ware (1984) extend this by treating the non-linear case, considering specifically the analysis of binary responses. An appreciation of these models in a broader and more theoretical context is conveyed in Breslow and Clayton (1993) and McCulloch (1997). In section 3.6 we discuss estimation for random effects models from an empirical Bayes perspective. Some further general comments are also

given there.

3.4 Two Simple Methods for Overdispersed Binomial Data

Well established methods exist for the regression analysis of binomial data, with the logistic-linear model being perhaps the best known and most popular. Extensions of this model are possible, to address the problem of extraneous variance in the data, known as extra-binomial variation (EBV) or more generally as overdispersion. EBV is manifest whenever the observed variability in the response proportions exceeds the nominal level predicted by the binomial model. Models for overdispersed binomial data have been proposed for example by Williams (1975, 1982) and Moore (1987), and an extension to the multinomial setting is described by O'Hara-Hines and Lawless (1993). We review two basic methods here, representing a likelihood and a quasi-likelihood approach.

3.4.1 Maximum Likelihood Estimation, Modified to Incorporate Extra-Binomial Variation

One explanation for EBV is based on the assumption that the correlation between any two observations within a given cluster is some non-zero value, say ρ. This is equivalent to assuming an exchangeable correlation structure for the data, in the sense that it implies a common correlation between any two observations within the same cluster.

Denoting individual responses as successes or failures, suppose that the ith cluster gives rise to R_i successes and $m_i - R_i$ failures. Assume further that associated with this cluster are explanatory variables $x_i = (x_{i1}, ..., x_{ip})^T$, $i = 1, ..., K$ where K is the number of clusters. Note that the following development only allows for covariates defined on the level of the cluster, so that the data can be expressed in collapsed form; that is, each

observation corresponds to the information from an entire cluster, namely the number of successes observed, the cluster size and corresponding covariate vector. This method would not be appropriate, for example, for modelling smoking rates as a function of gender, taking schools to be clusters, since gender is an individual-level covariate and generally not constant within schools.

The logistic model assumes that $R_i \sim \text{Bin}(m_i, \theta_i)$ where

$$\theta_i = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$$

is the probability of a success in cluster i, and $\beta = (\beta_1, ..., \beta_p)^T$ are the parameters to be estimated.

EBV may be modelled by assuming that R_i is binomially distributed only conditional on some value p_i, but that p_i itself is a realization of a random variable P_i supported on $[0, 1]$ with mean and variance depending on θ_i. Specifically,

$$\begin{aligned} R_i|P_i = p_i &\sim \text{Bin}(m_i, p_i) \\ \text{with} \quad \text{E}(P_i) &= \theta_i \quad \text{and} \\ \text{Var}(P_i) &= \rho\ \theta_i(1 - \theta_i). \end{aligned}$$

It follows that unconditionally

$$\text{E}(R_i) = \text{E}(\text{E}(R_i|P_i = p_i)) = \text{E}(m_i p_i) = m_i \theta_i \tag{3.4.1}$$

and

$$\begin{aligned} \text{Var}(R_i) &= \text{E}(\text{Var}(R_i|P_i = p_i)) + \text{Var}(\text{E}(R_i|P_i = p_i)) \\ &= \text{E}(m_i p_i(1 - p_i)) + \text{Var}(m_i p_i) \end{aligned}$$

$$
\begin{aligned}
&= m_i[\theta_i - \rho\ \theta_i(1-\theta_i) - \theta_i^2] + m_i^2\rho\ \theta_i(1-\theta_i) \\
&= m_i\theta_i(1-\theta_i)[1+\rho(m_i-1)]. \qquad (3.4.2)
\end{aligned}
$$

Hence $\mathrm{Var}(R_i) = v_i w_i^{-1}$ where $v_i = m_i\theta_i(1-\theta_i)$ (the nominal binomial variance), and $w_i^{-1} = 1 + \rho(m_i - 1)$, which is the variance inflation factor due to overdispersion. In this formulation ρ is interpreted as the intra-cluster correlation that is expected to exist between any two of the binary observations from a given cluster. (Note that we can write R_i as a sum of individual responses $\sum_j W_{ij}$, where $W_{ij} \sim \mathrm{Bin}(1,\theta_i), j = 1,...,m_i$ and $\mathrm{corr}(W_{ij}, W_{ik}) = \rho, j \neq k$). Observe that if $\rho = 1$, then $\mathrm{Var}(R_i) = m_i^2\theta_i(1-\theta_i)$, and if $\rho = 0$, $\mathrm{Var}(R_i) = m_i\theta_i(1-\theta_i)$. These results are exactly as expected; a value of $\rho = 1$ suggests perfect positive correlation between the observations in a cluster, implying that either $R_i = m_i$, with probability θ_i (the case that all $W_{ij} = 1$), or $R_i = 0$, with probability $1 - \theta_i$ (the case that all $W_{ij} = 0$). It is easily seen that in this situation that

$$
\begin{aligned}
\mathrm{Var}(R_i) &= \mathrm{E}(R_i^2) - \mathrm{E}(R_i)^2 \\
&= m_i^2\theta_i - (m_i\theta_i)^2 \\
&= m_i^2\theta_i(1-\theta_i).
\end{aligned}
$$

In contrast, $\rho = 0$ implies independence between observations, and the usual binomial variance $\mathrm{Var}(R_i) = m_i\theta_i(1-\theta_i)$ is obtained.

Some further interesting points: consider the variance of the proportion of successes in the ith cluster. This is

$$
\begin{aligned}
\mathrm{Var}\left(\frac{R_i}{m_i}\right) &= \frac{1}{m_i^2}\mathrm{Var}(R_i) \\
&= \frac{\theta_i(1-\theta_i)}{m_i} + \frac{m_i - 1}{m_i}\rho\ \theta_i(1-\theta_i).
\end{aligned}
$$

Written in the above form we see that this variance can be expressed as the sum of two

parts, the first constituting the binomial component and the second the extra-binomial component. Furthermore, $\text{Var}(R_i/m_i)$ can be written as the weighted average

$$\text{Var}(\frac{R_i}{m_i}) = \frac{\theta_i(1-\theta_i)}{m_i}(1-\rho) + \theta_i(1-\theta_i)\rho.$$

From this we observe the following: if $\rho = 0$

$$\text{Var}(\frac{R_i}{m_i}) = \frac{\theta_i(1-\theta_i)}{m_i},$$

which is the variance of the mean of m_i independent observations from the distribution $\text{Bin}(1, \theta_i)$. If $\rho = 1$

$$\text{Var}(\frac{R_i}{m_i}) = \theta_i(1-\theta_i),$$

which is the variance of a single observation from $\text{Bin}(1, \theta_i)$. No matter where in the unit interval $[0, 1]$ ρ lies, $\text{Var}(R_i/m_i)$ can always be interpreted as a weighted average of the variance of a single observation from $\text{Bin}(1, \theta_i)$ and the variance of the mean of m_i independent observations from $\text{Bin}(1, \theta_i)$. This highlights the notion that as the intra-cluster correlation ρ approaches 0, the information from the ith cluster increases to that contained in m_i independent binary observations, and as ρ approaches unity the information in cluster i decreases, as expected, to that contained in just a single observation.

Note that the interpretation of $\text{Var}(R_i/m_i)$ as a weighted average requires that $0 \leq \rho \leq 1$, and assuming that $\text{E}(P_i) = \theta_i$ and $\text{Var}(P_i) = \rho\ \theta_i(1-\theta_i)$ automatically restricts ρ to lie in this interval. To see this, observe that since $\text{Var}(P_i) \geq 0$ and $0 \leq \theta_i \leq 1$, we must have $\rho \geq 0$. Furthermore,

$$\begin{aligned} 0 \leq\ & P_i \leq 1 \\ \Longrightarrow\ & P_i^2 \leq P_i, \\ \text{so}\ & \text{E}(P_i^2) \leq \text{E}(P_i) = \theta_i, \end{aligned}$$

from which it follows that

$$\begin{aligned} \mathrm{Var}(P_i) &= \mathrm{E}(P_i^2) - \mathrm{E}(P_i)^2 \\ &\leq \mathrm{E}(P_i) - \theta_i^2 \\ &= \theta_i(1-\theta_i), \end{aligned}$$

implying that $\rho \leq 1$.

Maximum likelihood estimation is possible once a particular distribution for P_i is chosen. A widely used model for overdispersed binomial data is the beta-binomial, which assumes that P_i follows a beta distribution; see for example Williams (1975), Crowder (1978) and Moore (1987). Recall that if $P_i \sim \mathrm{beta}(a_i, b_i)$, then

$$\mathrm{E}(P_i) = \frac{a_i}{a_i + b_i} \quad \text{and} \quad \mathrm{Var}(P_i) = \frac{a_i b_i}{(1 + a_i + b_i)(a_i + b_i)^2}, \quad a_i, b_i > 0.$$

Therefore the variance of P_i has exactly the form $\rho\,\theta_i(1-\theta_i)$, with $\rho = (a_i + b_i + 1)^{-1}$ and $\theta_i = a_i/(a_i + b_i)$; (this is provided that $a_i + b_i$ is constant for all i, which may or may not be a reasonable assumption). Unconditionally R_i has a beta-binomial distribution, with the likelihood contribution from the ith cluster being

$$\frac{(a_i + r_i - 1)^{(r_i)}(b_i + m_i - r_i - 1)^{(m_i - r_i)}}{(a_i + b_i + m_i - 1)^{(m_i)}}, \tag{3.4.3}$$

having observed $R_i = r_i$ successes out of m_i trials. Here $n^{(r)}$ means 'n to r factors', i.e. $n^{(r)} = n(n-1)\ldots(n-r+1)$. We will revisit this model in Chapter 5 and provide some further details there.

Note that the above developments are only appropriate in discussing overdispersion; they cannot be applied to the much rarer phenomenon of underdispersion, in which case the intra-cluster correlation ρ is negative. Nevertheless the variance formula (3.4.2) in

itself *does* also accomodate such models, and it is possible in principle to obtain negative estimates of ρ when the data exhibit less than the nominal amount of variability expected under a particular model. It is also appealing that equation (3.4.2) yields the intuitively sensible result for the degenerate case $\rho = -1$. This describes a situation of perfect negative correlation between any two observations in a given cluster. It is clear that this can only arise if all clusters are of size two, each having one response equal to 1 and the other equal to 0. In other words, $R_i = 1$ with probability one for all i, implying that $\mathrm{Var}(R_i) = 0$; this is also the value given by formula (3.4.2) for $\rho = -1$ and $m_i = 2$.

3.4.2 Model Fitting Using Quasi-Likelihood

In the absence of EBV maximum likelihood estimates for $\boldsymbol{\beta}$ can be calculated by iterative reweighted least squares as described in Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). Given initial estimates $\boldsymbol{\beta}_{(0)}$, updated estimates $\boldsymbol{\beta}_{(1)}$ are given by

$$\boldsymbol{\beta}_{(1)} = (X^T V X)^{-1} X^T V Y \tag{3.4.4}$$

where X is the $K \times p$ design matrix of covariates, $V = \mathrm{diag}(v_i)$ and $Y = (Y_1, ..., Y_K)^T$ is the vector of adjusted response variables, with $Y_i = \sum_s x_{is}\beta_s + (R_i - m_i\theta_i)/v_i$. Note that in this notation $\mathrm{Var}(Y_i) = v_i^{-1}$, or equivalently, $\mathrm{Var}(Y) = V^{-1}$. Further, V and Y in (3.4.4) are evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_{(0)}$; because of the dependence of V and Y on $\boldsymbol{\beta}$, (3.4.4) must be solved iteratively to obtain the solution $\hat{\boldsymbol{\beta}}$.

Suppose now that overdispersion is suspected, but that no particular model for it is adopted. Thus the distribution of R_i is not fully specified. Direct maximum likelihood cannot be used in this case, but knowing the first two moments allows us to apply a quasi-likelihood method (Wedderburn, 1974). In this case we express the variance of Y as $\mathrm{Var}(Y) = (WV)^{-1}$ where $W = \mathrm{diag}(w_i)$, and the weighted least squares equations,

analogous to (3.4.4), are

$$\beta_{(1)} = (X^T W V X)^{-1} X^T W V Y. \tag{3.4.5}$$

The quasi-likelihood estimate $\hat{\beta}$ can be obtained from any logistic regression program that will allow a set of weights $\{w_i\}$ to be specified in the fitting procedure. However the weights in (3.4.5) depend on ρ which is initially unknown. Williams (1982) suggests an iterative procedure to estimate ρ. He notes that if the weights w_i were known and specified in the model, then the adjusted goodness-of-fit statistic

$$\chi^2 = \sum_{i=1}^{K} \frac{(R_i - m_i\theta_i)^2}{m_i\theta_i(1-\theta_i)w_i^{-1}} \tag{3.4.6}$$

(approximately the weighted sum of squares of residuals $(Y - X\hat{\beta})^T W V (Y - X\hat{\beta})$), would have asymptotically a chi-squared distribution on $K - p$ degrees of freedom. Now suppose for the moment that the m_i are equal, i.e. $m_i = m \; \forall \; i$, and a logistic model is fit without specifying any weights. Then from (3.4.2) and (3.4.6) it follows that in the presence of EBV the approximation

$$\mathrm{E}(\chi^2) = (K - p)[1 + \rho(m - 1)] \tag{3.4.7}$$

holds, implying that the heterogeneity factor $1 + \rho(m-1)$ can be estimated by $\chi^2/(K-p)$. Since in this simple case the weight matrix W is proportional to the identity matrix, the parameter estimates obtained from equations (3.4.4) and (3.4.5) are identical, but if (3.4.4) is used, the estimated covariance matrix of $\hat{\beta}$ must be scaled up by this heterogeneity factor. Equation (3.4.7) indicates that if a logistic model assuming no EBV is fit to data that are in fact overdispersed, the χ^2 goodness-of-fit statistic will tend to be systematically larger than $K - p$ on account of the overdispersion. For moderate to large cluster sizes, the

inflation factor multiplying $K-p$ can be rather significant, even for ρ close to 0. For instance, with clusters of size 50, an intra-cluster correlation as small as 0.02 will result in a two-fold inflation of the nominal χ^2 statistic. Hence, assuming that a p-parameter model is fit to data on K clusters, overdispersion is evident whenever $\chi^2 >> K-p$.

Relaxing the assumption of equal m_i, Williams gives two equations analogous to (3.4.7); the first assumes that χ^2 comes from a logistic fit with W the identity matrix, the second that χ^2 comes from a fit with prior weights already specified. These equations ((3.3) and (3.4) respectively in his paper) are

$$\mathrm{E}(\chi^2) = K-p+\rho\sum_i\{(m_i-1)(1-v_iq_i)\} \quad \text{if } W=I \tag{3.4.8}$$

$$\mathrm{E}(\chi^2) = \sum_i[w_i(1-w_iv_iq_i)\{1+\rho(m_i-1)\}] \quad \text{if } W\neq I, \tag{3.4.9}$$

where q_i is the ith diagonal element of $X(X^TWVX)X^T$. The iterative estimation scheme proceeds by equating the observed value of χ^2 to its expectation and solving for ρ. Thus one initially assumes that $\rho=0$, obtains χ^2 from the fit of the logistic model and compares this to the χ^2_{K-p} distribution. If χ^2 is unduly large, conclude that $\rho>0$ and calculate the current moment estimate $\hat{\rho}$ using (3.4.8), replacing $\mathrm{E}(\chi^2)$ with χ^2. Using weights $w_i=\{1+\hat{\rho}(m_i-1)\}^{-1}$ one then reestimates $\boldsymbol{\beta}$ using (3.4.5), recalculates χ^2 and the updated estimate of ρ using (3.4.9). Now one computes updated weights and repeats this process until χ^2 is sufficiently close to $K-p$, or equivalently, convergence in ρ has occurred.

We propose two equations similar in spirit to (3.4.8) and (3.4.9), which are however simpler and have more intuitive appeal. Returning again to the simplest case of equal cluster sizes, observe that the estimate of ρ from (3.4.7) is

$$\hat{\rho}=\frac{\chi^2-(K-p)}{(K-p)(m-1)}. \tag{3.4.10}$$

If K is large relative to p, this equation is roughly similar to a cruder approximation one

would obtain by replacing the denominator on the right side of (3.4.10) by $K(m-1)$. This allows us to generalize (3.4.10) immediately to the case of unequal cluster sizes: a reasonable estimate of ρ in this case is

$$\hat{\rho} = \frac{\chi^2 - (K-p)}{\sum(m_i - 1)}. \tag{3.4.11}$$

Finally, if χ^2 comes from a weighted model fit then an updated estimate of ρ must reflect the impact of the current weights w_i. In (3.4.11) it is assumed that each observation receives unit weight; if instead each observation receives weight w_i then it is sensible to replace K in equation (3.4.11) with $\sum w_i$ and $\sum(m_i - 1)$ with $\sum w_i(m_i - 1)$. Hence we obtain

$$\hat{\rho} = \frac{\chi^2 - (\sum w_i) + p}{\sum w_i(m_i - 1)}. \tag{3.4.12}$$

(Note that (3.4.12) reduces to (3.4.11) when $w_i = 1$ for all i). The iterative scheme proposed by Williams can be carried out by replacing the estimates from (3.4.8) and (3.4.9) with the simpler estimates (3.4.11) and (3.4.12); the results are identical and convergence using (3.4.11) and (3.4.12) is only slightly slower. Furthermore, using these alternative equations avoids the need of having to find v_i and q_i.

As an aside, a heuristic justification can be given for replacing K with $\sum w_i$ and similarly weighting the sum in the denominator of (3.4.12). With $\rho > 0$ each $w_i = \{1 + \hat{\rho}(m_i - 1)\}^{-1}$ is less than one, so $\sum w_i < K$. This suggests that one can perceive an analysis of K clusters exhibiting overdispersion as equivalent to an analysis of $\sum w_i < K$ clusters in which this has been adjusted for, i.e. showing no more extra variation. (We note here that in contrast to section 3.4.1, this quasi-likelihood approach also admits the added flexibility of allowing for underdipersion. Negative estimates of ρ are possible as long as $\hat{\rho} > -(m-1)^{-1}$ where $m = \max\{m_i\}$).

An attractive feature of this method is the fact that it can be implemented using standard software packages for generalized linear models, such as GLIM or SAS.

3.5 Generalized Estimating Equations

A versatile and widely utilized approach for the analysis of clustered data, which also accomodates covariates specific to each individual observation, is based on the use of generalized estimating equations (GEEs), as introduced formally by Liang and Zeger (1986) and Zeger and Liang (1986). This approach is an extension of generalized linear models to the analysis of longitudinally or otherwise clustered data. It involves the specification of estimating equations which give consistent estimates of regression parameters and their variances, while accounting for the correlation structure inherent in the clusters. The user must settle on an assumed form of this correlation structure in order to fit a GEE model, but the resulting estimates are robust to possible misspecification of the within-cluster dependence. Over the past 10 years a very large literature dealing with applications of the GEE methodology has developed. Some key references among many include Prentice (1988), who extended the GEE approach to allow joint estimation of the parameters in both the marginal means as well as the pairwise correlations (see also Zhao and Prentice (1990)); Zeger et al. (1988), who discuss applications for both population-averaged and subject-specific contexts; Lipsitz et. al (1994), who discuss issues of practical performance, and Fitzmaurice and Laird (1995), who describe applications to clustered bivariate responses.

3.5.1 Overview (IEE and GEE)

In this section we briefly explain how the GEE approach operates; the development closely parallels that given in Liang and Zeger (1986). The subsequent sections will describe a

variety of models which can be fit using this method.

As in section 3.2 let Y_{ij} denote the response of the jth individual in cluster i, $j = 1, ..., n_i$ and $i = 1, ..., K$; (in a longitudinal analysis Y_{ij} would refer to the jth observation on individual i). Associate with each Y_{ij} a vector of p explanatory covariates $x_{ij} = (x_{ij1}, ..., x_{ijp})^T$. Further, let $Y_i = (Y_{i1}, ..., Y_{in_i})^T$ represent the $n_i \times 1$ vector of responses from the ith cluster and $X_i = (x_{i1}, ..., x_{in_i})^T$ the associated $n_i \times p$ matrix of covariate values. We assume that Y_{ij} has a marginal density of the exponential family form

$$f(y_{ij}) = \exp\{(y_{ij}\theta_{ij} - a(\theta_{ij}) + b(y_{ij}))\phi\} \tag{3.5.1}$$

where $\theta_{ij} = h(\eta_{ij})$, $\eta_{ij} = x_{ij}'\beta$ and ϕ is a scale parameter. (The function $h(.)$ will vary with the choice of link function used to relate the random and systematic components of a particular model, say $\eta = g(\mu)$, where $\mu = \mathrm{E}(Y)$. Every distribution of the form (3.5.1) has a special link function for which θ, called the canonical parameter, equals the linear predictor η. These link functions are referred to as canonical links). It is easily shown that

$$\mu_{ij} = \mathrm{E}(Y_{ij}) = a'(\theta_{ij}) \quad \text{and} \quad v_{ij} = \mathrm{Var}(Y_{ij}) = a''(\theta_{ij})/\phi, \tag{3.5.2}$$

where a dash represents differentiation with respect to θ. Instead of modelling the joint distribution of the responses within a cluster, the GEE approach only makes an assumption about the first two moments of Y_i. In this sense it can be viewed as a multivariate analogue of quasi-likelihood (Wedderburn (1974), McCullagh (1983)).

Consider now the estimation of a set of regression parameters β_I, assuming to begin with that observations are independent. The familiar system of score equations one needs to solve in this case is the set of independence estimating equations (IEEs)

$$\sum_{i=1}^{K} X_i^T \Delta_i S_i = 0, \tag{3.5.3}$$

where $\Delta_i = \text{diag}(\partial\theta_{ij}/\partial\eta_{ij})$, $S_i = Y_i - \mu_i$ and $\mu_i = (\mu_{i1}, ..., \mu_{in_i})^T$. Note that when using the canonical link function, Δ_i is simply an $n_i \times n_i$ identity matrix. The estimator $\hat{\beta}_I$ is the solution to (3.5.3). Given that the particular model which was chosen (i.e. the particular form of (3.5.1)) is correct, the asymptotic covariance matrix of $\hat{\beta}_I$ (as $K \to \infty$) is given by

$$V_N(\hat{\beta}_I) = (\sum_{i=1}^{K} X_i^T \Delta_i A_i \Delta_i X_i)^{-1}, \tag{3.5.4}$$

where A_i is the $n_i \times n_i$ matrix $\text{diag}\{a''(\theta_{ij})\}$. This is often referred to as the naive or model-based variance estimate. Note that (3.5.4) depends on β_I through A_i and possibly Δ_i, and can be consistently estimated by replacing β_I with $\hat{\beta}_I$. A covariance matrix which is robust to model misspecification (and hence referred to as the robust estimate) is given by

$$V_R(\hat{\beta}_I) = V_N(\hat{\beta}_I) \{\sum_{i=1}^{K} X_i^T \Delta_i \text{Cov}(Y_i) \Delta_i X_i\} V_N(\hat{\beta}_I) \tag{3.5.5}$$

and is consistently estimated by substituting the moment estimate $(Y_i - \mu_i)(Y_i - \mu_i)^T$ for $\text{Cov}(Y_i)$, and evaluating this as well as A_i and Δ_i at $\hat{\beta}_I$. Both $\hat{\beta}_I$ and $\hat{V}_R(\hat{\beta}_I)$ are consistent given only that the regression model specified for $\text{E}(Y_{ij})$ is correct. The form of equation (3.5.5) follows from an argument based on Taylor series expansions, laid out carefully in Royall (1986). The implication of using $\hat{\beta}_I$ in conjunction with the robust variance estimate (3.5.5) is that asymptotically valid inferences will be drawn even if assumptions about the model are incorrect. In particular, (3.5.5) is robust to the misspecification of $\text{Cov}(Y_i)$ which results from assuming independence between observations within a cluster, i.e. not taking intra-cluster correlation into account. Hence the IEEs (3.5.3) may be recommended for practical purposes whenever the association between pairs of observations is considered of little interest in itself. Apart from this the cost of using the estimator $\hat{\beta}_I$ is in terms of a loss of efficiency; the stronger the intra-cluster correlation, the greater the loss of efficiency. The

GEE approach provides a way of obtaining regression estimates which are more efficient by making explicit assumptions about the correlation structure within clusters. At the same time the desirable property that variance estimates can be constructed which are robust to possible failure of these assumptions, is retained.

In order to write down the generalized estimating equations for a set of parameters β_G, analogous to (3.5.3), we define

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \phi, \tag{3.5.6}$$

which is the covariance matrix assumed for Y_i. V_i hinges on the $n_i \times n_i$ working correlation matrix $R_i(\alpha)$, which the user must specify; α here simply denotes a vector of parameters which fully specifies R_i. Note that if $R_i(\alpha)$ is in fact the true correlation matrix, then $V_i = \mathrm{Cov}(Y_i)$. The GEEs are defined to be

$$\sum_{i=1}^{K} D_i^T V_i^{-1} S_i = 0, \tag{3.5.7}$$

where $D_i = A_i \Delta_i X_i$. Note that these equations reduce to the score equations (3.5.3) if $R_i(\alpha)$ is the identity matrix. Let $\hat{\beta}_G$ be the solution to (3.5.7). Similar to (3.5.4) and (3.5.5), the model-based variance estimate is

$$V_N(\hat{\beta}_G) = (\sum_{i=1}^{K} D_i^T V_i^{-1} D_i)^{-1} \tag{3.5.8}$$

and the robust estimate is

$$V_R(\hat{\beta}_G) = V_N(\hat{\beta}_G) \{\sum_{i=1}^{K} D_i^T V_i^{-1} \mathrm{Cov}(Y_i) V_i^{-1} D_i\} V_N(\hat{\beta}_G). \tag{3.5.9}$$

Estimation of β_G is carried out by a procedure equivalent to iteratively reweighted least squares. At each iteration, updated moment estimates of α and possibly ϕ are calculated,

based on the current estimate of β_G.

We shall discuss specific choices for the correlation matrix in the next section, and only mention here that α is typically a function of the estimated Pearson residuals

$$\hat{r}_{ij} = \{y_{ij} - \hat{\mu}_{ij}\}/\{\phi \hat{v}_{ij}\}^{1/2},$$

where $\hat{\mu}_{ij} = \mu_{ij}(\hat{\beta}_G)$ and $\hat{v}_{ij} = v_{ij}(\hat{\beta}_G)$; for models in which the scale parameter ϕ is not equal to one, it can be estimated by

$$\hat{\phi}^{-1} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \hat{r}_{ij}^2 / (\sum_{i=1}^{K} n_i - p).$$

3.5.2 Marginal Models and Possible Correlation Structures

A variety of models with differing interpretations can be fit using GEE, some of which can be classified as marginal models; we shall discuss these first.

In a marginal regression model the marginal expectation of the response is modelled as a function of explanatory variables. If multiple observations are made on a number of individuals, as in a longitudinal study, the GEE approach will separately model the correlation we might expect between responses from the same individual; the same holds if study units consist of clusters of a general nature, each of which give rise to several observations. Estimation of regression coefficients and correlation parameters is typically achieved using one set of estimating equations of the form (3.5.7), although Prentice (1988) describes a more involved approach in which both response probabilities and correlations are modelled as functions of explanatory variables. This requires two sets of estimating equations, one for the mean specification of the model and the other for the correlation parameters α. We shall discuss this approach further in section 3.5.5.

A wide variety of choices for $R(\alpha)$ can be specified using GEE. Existing SAS macros, distributed for example by M. R. Karim (©1989, Department of Biostatistics, The

Johns Hopkins University) and U. Groemping (©1993, Fachbereich Statistik, Universität Dortmund, Germany), admit independent, exchangeable, m-dependent, user-specified and unspecified correlation structures. (More recently the fitting of GEE models has become possible using the procedure GENMOD in SAS Version 6.12).

The independent correlation structure assumes that $R(\alpha)$ is the identity matrix; a GEE fit under this assumption is equivalent to an analysis assuming independence between all observations.

Exchangeable correlation may be a reasonable postulation for cross-sectional clusters in which observations are not ordered in any special manner. The assumption is that a common correlation ρ is shared between any two observations within the same cluster, i.e. $\text{Corr}(Y_{ij}, Y_{ik}) = \rho,\ j \neq k$. (In this case $\dim(\alpha) = 1$, with $\alpha = \rho$). The parameter ρ, and more generally α when dealing with other correlation structures, is estimated by borrowing strength across the K clusters. In this case ρ is estimated by

$$\hat{\rho} = \hat{\phi}\,[\sum_{i=1}^{K}\sum_{j>j'}\hat{r}_{ij}\hat{r}_{ij'}]/[\sum_{i=1}^{K}\frac{1}{2}n_i(n_i-1)-p]. \tag{3.5.10}$$

The m-dependent correlation structure is especially suitable for longitudinal data in that it allows the correlation between repeated observations to depend on the lag between the observations; it assumes that

$$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} \rho_{k-j} & \text{for } k = j+1, ..., j+m \\ 0 & \text{for } k > j+m. \end{cases}$$

Similar formulae as for the case of exchangeable correlation, based on functions of residuals, exist for the estimation of ρ_{k-j}; see Liang and Zeger (1986).

If one is dealing with small cluster sizes and has strong prior convictions about the intra-cluster correlation, the user-specified correlation structure may be the most appealing

option. In this case $R(\alpha)$ is not estimated but provided a *priori* by the user. In contrast, one may wish to leave $R(\alpha)$ totally unspecified and have GEE estimate it by the sample correlation matrix based on the data. This requires equi-sized clusters, and since $\frac{1}{2}n(n-1)$ parameters need to be estimated, this choice is only sensible for small cluster sizes.

Note that variable cluster sizes are admissible only under the assumption of independent or exchangeable correlation. All other structures require equi-sized clusters.

The software currently available to fit GEE models has the limitation that once a correlation structure is specified, that same structure is assumed to apply to all clusters in the data set. On occasion it would desirable to relax this condition. Suppose, for example, that we find an exchangeable correlation structure to be adequate, but instead of an estimate of intra-cluster correlation that applies to all clusters, we identify groups of clusters for each of which we would like a separate estimate of the exchangeable correlation. Examining the WSPP3 data for instance, we might wish to analyze smoking rates in grades 7 and 8, and define cross-sectional clusters as groups of students in the same school and the same grade. This would give rise to two logical groups of clusters, one for the grade 7 responses and one for the grade 8 responses. Hence we might be interested in estimating a separate correlation parameter for each of these groups. Since estimation of a single correlation parameter proceeds by borrowing strength across all clusters, estimation of several such parameters could logically be achieved by borrowing strength only across the clusters which define the particular groups that we are interested in differentiating. For the example above, suppose that C_1 and C_2 denote the set of clusters of grade 7 and grade 8 responses respectively. Denoting the intra-cluster correlations in the two groups as ρ_1 and ρ_2, the estimate of ρ_k would then be, similar to equation (3.5.10),

$$\hat{\rho}_k = \hat{\phi}\,[\sum_{i\in C_k}\sum_{j>j'}\hat{r}_{ij}\hat{r}_{ij'}]/[\sum_{i\in C_k}\frac{1}{2}n_i(n_i-1)-p],\quad k=1,2. \tag{3.5.11}$$

One question which arises here concerns the minimum number of clusters one should have

in any given group.

A second and similar generalization with respect to modelling intra-cluster correlation involves estimation of separate correlation parameters for subgroups of observations within clusters. This achieves a compromise between the exchangeable and unspecified correlation structures. Suppose observations within clusters can be further grouped on some basis, (say the gender of students within schools), and it is of interest to estimate the correlation between individuals in the same group within a cluster, as well as the correlation between individuals in different groups within a cluster. Using the example of gender distinction, three correlation parameters (ρ_f, ρ_m and ρ_{fm}) could be estimated, by borrowing strength across all clusters but restricting calculations first to only females in the clusters, then to males, and finally to the mixed gender pairs. In this case

$$\hat{\rho}_f = \hat{\phi}\,[\sum_{i=1}^{K} \sum_{j>j',j,j' \in F_i} \hat{r}_{ij}\hat{r}_{ij'}]/[\sum_{i=1}^{K} \frac{1}{2} n_{fi}(n_{fi}-1) - p], \tag{3.5.12}$$

where F_i is the set of females and n_{fi} the number of females in cluster i. A similar estimate can be constructed for ρ_m, by replacing F_i and n_{fi} in (3.5.12) by the analogous quantities M_i and n_{mi}. Finally,

$$\hat{\rho}_{fm} = \hat{\phi}\,[\sum_{i=1}^{K} \sum_{j \in F_i, j' \in M_i} \hat{r}_{ij}\hat{r}_{ij'}]/[\sum_{i=1}^{K} n_{fi}n_{mi} - p]. \tag{3.5.13}$$

3.5.3 Survival Models

In the framework of dichotomous responses there are applications in which we are not interested so much in the marginal probability of response as in an analysis of the data from a time-to-event or current status perspective. With the WSPP3 data for example, we might consider modelling the probability of failure (having smoked at least once) by time t, given survival (having never smoked) until time $t-1$. Examining data for grades 6, 7 and 8

for instance, such an analysis would include only non-smokers at grade 6 (whose responses would be smoking status in grade 7) and non-smokers at grade 7 (whose responses would be smoking status in grade 8). A grouped proportional hazards model (Cox, (1972), Cox and Oakes (1984)) is appropriate for this type of modelling, and can be combined with the GEE methodology to account for the clustering due to schools. One needs to specify a model with binomial error structure in conjuction with the complementary log-log link.

To examine this idea more closely, suppose that observations are made at time points t_r, $r = 1, 2, 3, \ldots$. We are interested in modelling the probability of failure during the rth time interval, given survival up until time t_{r-1}. To this end, define $Y_{ij}(r)$ as

$$Y_{ij}(r) = \begin{cases} 1 & \text{if } T_{ij} \leq t_r, \text{ given } T_{ij} > t_{r-1} \\ 0 & \text{if } T_{ij} > t_r \end{cases}$$

where T_{ij} denotes the survival time (time to first smoking) of individual j in cluster i. We wish to construct a model for

$$\begin{aligned} p_{ij}(r) &= \mathrm{E}(Y_{ij}(r)) \\ &= P(t_{r-1} < T_{ij} \leq t_r | T_{ij} > t_{r-1}), \quad r = 1, 2, 3, \ldots . \end{aligned} \tag{3.5.14}$$

Suppose now that the hazard function of the distribution governing the 'lifetimes' T_{ij} has the Cox proportional hazards form (Cox, (1972))

$$h(t; x_{ij}) = h_0(t) e^{x_{ij}' \beta},$$

where x_{ij} is a vector of time-independent covariates with associated coefficients β, and $h_0(t)$ is a baseline hazard function not depending on x_{ij}. The survivor function of T_{ij} can

be written as

$$\begin{aligned} S_{ij}(t) = P(T_{ij} > t) &= \exp\{-\int_0^t h(u; x_{ij})du\} \\ &= \exp\{-e^{x_{ij}'\beta}\int_0^t h_0(u)du\}. \end{aligned}$$

Therefore

$$\begin{aligned} p_{ij}(r) &= \frac{S_{ij}(t_{r-1}) - S_{ij}(t_r)}{S_{ij}(t_{r-1})} \\ &= 1 - \exp\{-e^{x_{ij}'\beta}[\int_0^{t_r} h_0(u)du - \int_0^{t_{r-1}} h_0(u)du]\} \\ &= 1 - \exp\{-e^{g(t_r)+x_{ij}'\beta}\}, \end{aligned}$$

where $g(t_r) = \log \int_{t_{r-1}}^{t_r} h_0(u)du$, so that

$$\log\{-\log(1 - p_{ij}(r))\} = g(t_r) + x_{ij}'\beta,$$

implying the complementary log-log link. The simplest choice for $g(t_r)$ is a linear function of t_r. Thus for example if we are analyzing grade 7 and 8 responses we might choose $g(t_r) = \beta_0 + \beta_1 t_r$ where $t_r = 0$ for a grade 7 observation (say $r = 1$), and $t_r = 1$ for a grade 8 observation (say $r = 2$). The extension to a larger number of time points is straightforward; in general, if responses at ℓ time points are to be analyzed, $g(t_r)$ can be written as a linear combination of $\ell - 1$ indicator variables plus a constant term, $r = 1, ..., \ell$.

Depending upon the application and the analysis of interest, such survival models in combination with the GEE methodology can be a useful means of analyzing data that are both longitudinally and cross-sectionally clustered. In the WSPP3 data for instance, multiple responses over time can be examined using the approach described above, while at the same time accounting for the clustering induced by the different schools, using GEE. The estimating equations are exactly of the form (3.5.7); the only change necessary

when using the complementary log-log link as opposed to the canonical logistic link is in terms of the matrix Δ_i, which will no longer be the identity matrix. Recall that $\Delta_i = \text{diag}(\partial\theta_{ij}/\partial\eta_{ij})$. For binary data $Y_{ij}(r)$, $\theta_{ij} = \log\{p_{ij}(r)/(1-p_{ij}(r))\}$, and letting $\eta_{ij} = g(t_r) + x'_{ij}\beta$ it follows that

$$\theta_{ij} = \log[\exp\{e^{\eta_{ij}}\} - 1], \tag{3.5.15}$$

and hence

$$\frac{\partial\theta_{ij}}{\partial\eta_{ij}} = \frac{\exp\{\eta_{ij} + e^{\eta_{ij}}\}}{\exp\{e^{\eta_{ij}}\} - 1}. \tag{3.5.16}$$

Thus for the complementary log-log link,

$$\Delta_i = \text{diag}(\frac{\exp\{\eta_{ij} + e^{\eta_{ij}}\}}{\exp\{e^{\eta_{ij}}\} - 1}).$$

(Similar straightforward calculations will yield Δ_i for any other link function which might be of interest).

3.5.4 Transition Models

Transition models also constitute a useful method for analyzing longitudinal data. These models distinguish themselves from survival models in that they do not focus on the time to a certain event, but allow for transitions into and out of a set of states defined by the response variable. Suppose that multiple observations are made over time on each of a number of individuals. In contrast to marginal models, transition models describe the expectation of the response variable conditional on previous responses as well as covariate information. Let Y_{ijt} be the response of individual j in cluster i at time point t, $t = 1, ..., T$. Perhaps the most widely used transition models are Markov models, which assume that

the conditional distribution of Y_{ijt} given all previous responses depends only on the m previous responses $Y_{ijt-1}, ..., Y_{ijt-m}$, for some value of m. The observations on a particular subject are assumed to be conditionally independent. Thus, assuming for the moment that observations are independent between and within clusters, the likelihood contribution of individual j in cluster i for a Markov model of order $m = 1$ is

$$L_{ij}(y_{ij1}, ..., y_{ijT}) = f(y_{ij1}) \prod_{t=2}^{T} f(y_{ijt}|y_{ijt-1}),$$

with the full likelihood given by $\prod_{ij} L_{ij}$. Any software that fits GLMs can be used to fit such a model, by including Y_{ijt-1} in the set of predictor variables for observation Y_{ijt}. If observations within (cross-sectional) clusters are not independent, GEE can be used to account for this in a similar fashion as for the survival models of section 3.5.3.

For binary responses we can distinguish between two types of models. The first is the one discussed above and assumes that the effect of the covariates on the response Y_{ijt} is the same for both positive and negative responders at Y_{ijt-1}, after adjusting for Y_{ijt-1}. Such models simply include Y_{ijt-1} as an explanatory variable in the linear predictor for Y_{ijt}; they take the form

$$\text{logit}\, P(Y_{ijt} = 1|Y_{ijt-1} = y_{ijt-1}) = x_{ij}'\beta + \gamma y_{ijt-1}.$$

With reference to the WSPP3 data, if Y_{ijt} is an indicator of smoking status for individual j in school i at time t, the odds of this individual smoking at time t, having adjusted for the effects of the covariates contained in x_{ijt}, are e^{γ} times larger if that student was smoking at time $t-1$ as opposed to not smoking.

The second type of model assumes that the effect of the covariates on the response Y_{ijt} is different for positive and negative responders at Y_{ijt-1}. If this assumption holds one might either fit two separate logistic models to Y_{ijt} (one to the portion of the data with $Y_{ijt-1} = 0$, and one to the part with $Y_{ijt-1} = 1$), or one may combine these two into a single

model by including as predictors the previous response Y_{ijt-1} as well as the interaction of Y_{ijt-1} with each of the covariates under consideration. For note that the two models

$$\text{logit}\, P(Y_{ijt} = 1|Y_{ijt-1} = 0) = x'_{ijt}\beta_0$$
$$\text{logit}\, P(Y_{ijt} = 1|Y_{ijt-1} = 1) = x'_{ijt}\beta_1$$

are equivalent to

$$\text{logit}\, P(Y_{ijt} = 1|Y_{ijt-1} = y_{ijt-1}) = x'_{ijt}\beta_0 + y_{ijt-1}x'_{ijt}\gamma$$

where $\beta_1 = \beta_0 + \gamma$ (Diggle et al. (1994), Ch. 10). Continuing with the smoking example, γ is a measure of the covariate effect *differential* inherent in x_{ijt} when modelling smoking rates at time t and comparing smokers at time $t-1$ with non-smokers. In this case the odds of individual (i,j) smoking at time t, having adjusted for the effects the covariates, are $e^{x'_{ijt}\gamma}$ times larger if he or she was smoking at time $t-1$ as opposed to not smoking; note that this factor now depends on x_{ijt}. If $x_{ijt} = (1, x_{ijt1}, ..., x_{ijts})'$ and $\gamma = (\gamma_0, \gamma_1, ..., \gamma_s)'$, then a test of the hypothesis that $\gamma = (\gamma_0, 0, ..., 0)'$ is equivalent to testing whether or not the covariates have the same effect on the response Y_{ijt} regardless of the value of Y_{ijt-1} (i.e. whether or not the simple Markov model described above is adequate).

A logical extension can be made to include Y_{ijt-1} and Y_{ijt-2} in the model for Y_{ijt}. Alternatively it may be of interest to condition on the previous response as well as a dichotomous covariate. For example, one might wish to model smoking rates at a particular time point, adjusting for smoking status at the previous time point and the gender of the individual. Letting $Z_{ij} = 0$ for a male and 1 for a female, say, the four models of interest

$$\text{logit}\, P(Y_{ijt} = 1|Y_{ijt-1} = 0, Z_{ij} = 0) = x'_{ijt}\beta_{00}$$
$$\text{logit}\, P(Y_{ijt} = 1|Y_{ijt-1} = 0, Z_{ij} = 1) = x'_{ijt}\beta_{01}$$

$$\text{logit}\, P(Y_{ijt} = 1 | Y_{ijt-1} = 1, Z_{ij} = 0) = x'_{ijt}\beta_{10}$$
$$\text{logit}\, P(Y_{ijt} = 1 | Y_{ijt-1} = 1, Z_{ij} = 1) = x'_{ijt}\beta_{11}$$

can be combined into the following single model:

$$\text{logit}\, P(Y_{ijt} = 1 | Y_{ijt-1} = y_{ijt-1}, Z_{ij} = z_{ij}) =$$
$$x'_{ijt}\beta + y_{ijt-1}x'_{ijt}\gamma_1 + z_{ij}x'_{ijt}\gamma_2 + y_{ijt-1}z_{ij}x'_{ijt}\gamma_3.$$

Thus $\beta_{00} = \beta$, $\beta_{01} = \beta + \gamma_2$, $\beta_{10} = \beta + \gamma_1$ and $\beta_{11} = \beta + \gamma_1 + \gamma_2 + \gamma_3$. Now a test of $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$ would determine whether or not the covariates x_{ijt} have the same effect on the response probability regardless of the value of (Y_{ijt-1}, Z_{ij}). Similarly, a test of $H_0 : \gamma_2 = \gamma_3 = 0$ would determine whether or not the covariates have the same effect on the response regardless of an individual's gender (Z_{ij}). Testing $H_0 : \gamma_1 = \gamma_3 = 0$ would enable one to draw conclusions in a similar fashion about previous smoking status Y_{ijt-1}. As already indicated above, a more refined version of these hypothesis tests might only test that all components except the intercept term in γ_r are zero. This would allow the conditioning variables to affect the baseline response probability without impacting on the additional effect of the covariates.

Transition models are very useful for conditional analyses of longitudinal data, especially when strong dependence is suspected between successive observations on an individual. This is the case for instance in the WSPP3 data; a student's smoking status at time t can be shown to depend heavily on smoking status at time $t-1$ and even $t-2$; see Chapter 7 for further details.

Survival and transition models were discussed in the context of GEE here, but one should note that a random effects approach (see for instance section 3.6) can also be applied to such models as an alternative cluster-specific method of adjusting for extraneous school-to-school variability.

3.5.5 Symmetric Modelling of Mean and Correlation

Prentice (1988) describes an extension of the GEE approach for binary regression to allow estimation of parameters not only in a model for the response probabilities, but also in a similar model for the correlation parameters. Explicit modelling of the pairwise correlations in this manner is an alternative to the simpler methods of section 3.5.2 and may further improve the efficiency of the estimator $\boldsymbol{\beta}$ for the response rates. Instead of basing the estimate of the correlation parameter $\boldsymbol{\alpha}$ on functions of standardized residuals, Prentice suggests a GEE estimator of $\boldsymbol{\alpha}$ which is constructed in much the same manner as the estimator for $\boldsymbol{\beta}$. He derives a second set of estimating equations for $\boldsymbol{\alpha}$, as follows:

Let $\mathrm{E}(Y_{ij}) = \mu_{ij} = p_{ij}$, $q_{ij} = 1 - p_{ij}$, $\boldsymbol{p}_i = (p_{i1}, ..., p_{in_i})$ and $D_i = A_i \Delta_i X_i$. Note that D_i is the $n_i \times p$ matrix of derivatives $\partial \boldsymbol{p}_i / \partial \boldsymbol{\beta}$, where $p = \dim(\boldsymbol{\beta})$. In this notation the generalized estimating equations (3.5.7) can be written as

$$\sum_{i=1}^{K} D_i^T V_i^{-1} (Y_i - \boldsymbol{p}_i) = 0. \tag{3.5.17}$$

To obtain (3.5.17) we needed to make certain assumptions about $\boldsymbol{p}_i$ and V_i, the mean and variance of Y_i. Hence in order to derive an analogous set of estimating equations for $\boldsymbol{\alpha}$ we need to make similar assumptions about the mean and variance of Z_i, the $\binom{n_i}{2}$ vector of sample correlations between the observations in cluster i. As in Prentice's paper we let $Z_i = (Z_{i12}, ..., Z_{i1n_i}, Z_{i23}, ..., Z_{in_i-1n_i})'$, where Z_{ijk} is the sample correlation between observation j and k in cluster i, equal to

$$Z_{ijk} = Z_{ijk}(\boldsymbol{\beta}) = \frac{(Y_{ij} - p_{ij})(Y_{ik} - p_{ik})}{(p_{ij} q_{ij} p_{ik} q_{ik})^{1/2}}. \tag{3.5.18}$$

Letting $\mathrm{E}(Z_{ijk}) = \rho_{ijk}$ it follows that

$$\mathrm{Var}(Z_{ijk}) = w_{ijk} = 1 + (1 - 2p_{ij})(1 - 2p_{ik})(p_{ij} q_{ij} p_{ik} q_{ik})^{-1/2} \rho_{ijk} - \rho_{ijk}^2. \tag{3.5.19}$$

To show this we note that $\mathrm{Var}(Z_{ijk}) = \mathrm{E}(Z_{ijk}^2) - \rho_{ijk}^2$ and that

$$\begin{aligned}
\mathrm{E}(Z_{ijk}^2) &= (p_{ij}q_{ij}p_{ik}q_{ik})^{-1}\mathrm{E}\{(Y_{ij}Y_{ik} - Y_{ij}p_{ik} - Y_{ik}p_{ij} + p_{ij}p_{ik})^2\} \\
&= c\ \mathrm{E}\{Y_{ij}^2Y_{ik}^2 - 2Y_{ij}^2Y_{ik}p_{ik} + Y_{ij}^2p_{ik}^2 \\
&\quad -2(Y_{ij}Y_{ik}^2p_{ij} - 2Y_{ij}Y_{ik}p_{ij}p_{ik} + Y_{ij}p_{ij}p_{ik}^2) \\
&\quad +Y_{ik}^2p_{ij}^2 - 2Y_{ik}p_{ik}p_{ij}^2 + p_{ik}^2p_{ij}^2\}, \qquad (3.5.20)
\end{aligned}$$

where $c = (p_{ij}q_{ij}p_{ik}q_{ik})^{-1}$. Now since Y_{ij} is a 0-1 response, $Y_{ij}^r = Y_{ij}$, $r = 1, 2, \ldots$. Thus

$$\begin{aligned}
\mathrm{E}(Y_{ij}^rY_{ik}^s) &= \mathrm{E}(Y_{ij}Y_{ik}) \\
&= \mathrm{Cov}(Y_{ij}Y_{ik}) + p_{ij}p_{ik} \\
&= \rho_{ijk}(p_{ij}q_{ij}p_{ik}q_{ik})^{1/2} + p_{ij}p_{ik}
\end{aligned}$$

for $r, s = 1, 2, \ldots$. Using this result to carry through the expectation in equation (3.5.20) yields, upon some simplification and subtraction of ρ_{ijk}^2, equation (3.5.19).

Let $\rho_i = (\rho_{i12}, \ldots, \rho_{i1n_i}, \rho_{i23}, \ldots, \rho_{in_i-1n_i})'$ and $W_i = \mathrm{diag}\{w_{i12}, \ldots, w_{i1n_i}, w_{i23}, \ldots, w_{in_i-1n_i}\}$. Furthermore, let E_i be the $\binom{n_i}{2} \times r$ matrix of derivatives $\partial\rho_i/\partial\alpha$, where $r = \dim(\alpha)$. The estimating equations for α are, analogous to (3.5.17),

$$\sum_{i=1}^{K} E_i^T W_i^{-1}(Z_i - \rho_i) = 0. \qquad (3.5.21)$$

This set of equations allows one to estimate parameters in a flexible model for ρ_{ijk}. One possible choice might be $\rho_{ijk} = \alpha_0 + \alpha_1 z_{ij} + \alpha_2 z_{ik} + \alpha_3 z_{ij}z_{ik} + \alpha_4 n_i$, where z_{ij} indicates the gender of the jth individual in cluster i; this would assume that the correlation between two individuals in a given cluster depends on their gender as well as the cluster size.

One constraint to bear in mind when modelling ρ_{ijk} is the fact that, barring further restrictions, this coefficient must in any case lie in the interval $(-1, 1)$. By way of analogy,

one of the key reasons for choosing a logistic model for $E(Y_{ij}) = p_{ij}$ is to guarantee a fitted value $\hat{p}_{ij}$ which lies in $(0,1)$, while at the same time removing all constraints in estimating the parameters β. A similar provision would be useful when building a model for ρ_{ijk}.

Note in contrast to the estimating equations (3.5.17) that no working correlation matrix (other than the implicit identity matrix) is specified in (3.5.21). If the sizes of W_i, $i = 1, ..., K$ are small enough, a generalization to include a non-trivial working correlation matrix is straightforward. Note however that for sizeable clusters, W_i, of dimension $\binom{n_i}{2} \times \binom{n_i}{2}$, is prohibitively large. For example, a cluster of size 100 would necessitate a 4950×4950 matrix W_i. In such cases the diagonal structure of this matrix should be exploited to avoid having to define it explicitly at all. If W_i is diagonal, $E_i^T W_i^{-1}$ does not have to be constructed internally through matrix multiplication but can be defined at once as a single matrix, say T_i. Let $\nu_{i1}, ..., \nu_{ir}$ denote the rows of E_i^T, and let $\omega_i = \text{diag}(W_i^{-1})$. The rows of T_i are simply $\nu_{i1} \times \omega_i, ..., \nu_{ir} \times \omega_i$, where '$\times$' indicates element-wise multiplication. Thus the computational burden is reduced from specifying W_i to defining the $r \times \binom{n_i}{2}$ matrix T_i, which is more manageable.

The joint estimation of (β, α) proceeds by iterating back and forth between the two sets of estimating equations (3.5.17) and (3.5.21) until convergence is achieved at a value $(\hat{\beta}(\hat{\alpha}), \hat{\alpha}(\hat{\beta}))$. Naive and robust variance estimates for α can be constructed in exactly the same fashion as for β (refer to equations (3.5.8) and (3.5.9)). More generally however one will expect nonzero covariances between the elements of β and α. In this case a covariance estimate of the joint parameter estimate $(\hat{\beta}, \hat{\alpha})$ is available and is given by expression (15) in Prentice (1988).

3.5.6 Application of GEE to Random Effects Models

As a final note in this expository section on generalized estimating equations, we briefly highlight the work of Waclawiw and Liang (1993, 1994), which provides a good example

of how the GEE methodology may also be used to draw cluster-specific inferences in the context of a random effects model. They propose a cyclical algorithm based on the use of estimating equations to fit such a model, obtaining estimates of the fixed effects β, the random effects, as well as the variance of the random effects distribution. A cycle of three steps is iterated repeatedly until convergence is achieved in one of the parameters of interest (usually the estimate of the random effects variance).

We shall describe the procedure for the specific case of the logistic - univariate normal model as discussed in Waclawiw and Liang (1994) and in the next section. In principle however, it can be adapted to other error distributions in the exponential family, other link functions and different choices for the random effects distribution.

Begin by assuming that

$$\begin{aligned} Y_{ij}|x_{ij}, b_i &\sim \text{Bin}(1, p_{ij}), \qquad \text{where} \\ \theta_{ij} &= \log(\frac{p_{ij}}{1-p_{ij}}) = x'_{ij}\beta + b_i, \\ b_i &\overset{\text{iid}}{\sim} N(0, \sigma^2). \end{aligned} \tag{3.5.22}$$

This model adjusts for the heterogeneity across clusters by assuming the presence of an unobserved random effect for each cluster, conditional upon which observations are treated as independent. The random effects themselves are assumed to obtain from a Gaussian distribution with zero mean. See the next section for further details.

Starting with an assumed initial value for σ^2, the variance of the random effects, the GEE approach is used to estimate β, the vector of fixed effects. In order to implement GEE an expression for the marginal mean and variance of the response Y_{ij} is required. Under the above model formulation no closed form expression for $\text{E}(Y_{ij})$ exists. However

the cumulative Gaussian approximation to the logistic function,

$$\text{logit } \mathrm{E}(Y_{ij}) \;=\; \text{logit } \mathrm{E}(\mathrm{E}(Y_{ij}|b_i)) \;\simeq\; \frac{x_{ij}'\beta}{\sqrt{1+c^2\sigma^2}}, \qquad c=\frac{16\sqrt{3}}{15\pi}, \tag{3.5.23}$$

as cited in Zeger et al. (1988) and Johnson and Kotz (1970), is remarkably accurate over a reasonable range for the linear predictor and the variance σ^2 (say, $-3 \le x_{ij}'\beta \le 3$ and $\sigma^2 \le 2$). Zeger et al. (1988) further give an approximation for the marginal covariance matrix of the vector of responses from the ith cluster, based on a Taylor series expansion of the link function about $b_i = 0$. This is justified in that only an approximation for the covariance matrix is required to obtain consistent and reasonably efficient estimates for β using the GEE approach. Note that in this step of the procedure only an estimate of σ^2 is required; it does not actually involve the random effects themselves.

Having obtained an estimate of the fixed effects β, one can use these as well as the current estimate of the random effects variance to obtain estimates of the random effects b_i, $i = 1, ..., K$. These are computed as the solutions to estimating equations of the form

$$g_i(b_i, \beta, \sigma^2) = \sum_{j=1}^{n_i}(\hat{c_{ij}}y_{ij} + \hat{d_i} - \mathrm{E}(Y_{ij}|b_i)) = 0, \qquad i = 1, ..., K. \tag{3.5.24}$$

Optimal values of the shrinkage factors c_{ij} and d_i can be computed by minimizing the Godambe risk function (Godambe (1960)). This leads to a set of Stein-type estimating equations, each of the form

$$\sum_{j=1}^{n_i}(w_i y_{ij} + (1-w_i)\mathrm{E}(Y_{ij})) = \sum_{j=1}^{n_i}(\mathrm{E}(Y_{ij}|b_i)), \tag{3.5.25}$$

where the w_i are estimated weights ($0 < w_i < 1$) which will be functions of β and σ^2 (see also Liang and Waclawiw (1990)).

Finally, from the estimates computed for β and b_i, $i = 1, ..., K$, one can calculate an

updated estimate of σ^2. Note that

$$\mathrm{Var}(b_i) = \mathrm{E}(b_i^2) = \mathrm{E}(\hat{b_i^2}) - 2\mathrm{E}((\hat{b_i} - b_i)\hat{b_i}) + \mathrm{E}(\hat{b_i} - b_i)^2. \tag{3.5.26}$$

Waclawiw and Liang suggest estimating $\mathrm{Var}(b_i)$ by approximating the first term on the right hand side of equation (3.5.26) by the moment estimator $\sum_{i=1}^{K} \hat{b_i^2}/K$, assuming the cross-product term to be negligible and estimating the last term with the estimating function based estimator

$$\mathrm{E}(\hat{b_i} - b_i)^2 = \mathrm{E}\{g_i(\hat{b_i}, \beta, \sigma^2)/\frac{\partial g_i(\hat{b_i}, \beta, \sigma^2)}{\partial b_i}\}^2, \tag{3.5.27}$$

which is derived from the linear Taylor series approximation to $g_i(\hat{b_i}, \beta, \sigma^2) = 0$, expanding about b_i. Alternatively, this term could be estimated by the asymptotic posterior variance

$$\begin{aligned} \mathrm{Var}(b_i|y_i) &\simeq -(\frac{\partial^2 \log f(b_i|y_i)}{\partial b_i^2})^{-1} \\ &\simeq \frac{1}{K}\sum_{i=1}^{K}[\sum_{j=1}^{n_i} p_{ij}(1-p_{ij}) + \frac{1}{\sigma^2}]^{-1}, \end{aligned} \tag{3.5.28}$$

where $f(b_i|y_i)$ is the posterior distribution of b_i given the data from the ith cluster; note that this variance estimate also borrows strength across clusters by averaging over $i = 1, ..., K$.

With the updated estimate of σ^2 one can begin another cycle of the procedure and estimate the updated value of β, hence the random effects b_i, and finally a subsequent estimate of σ^2. Iteration in this fashion continues until convergence in one of the parameters, such as σ^2, is achieved.

Since the random effects variance is updated empirically after each iteration, this approach can be viewed as a combination of GEE and empirical Bayes estimation for random effects models. More will follow on this in next section and in Chapter 4, where these two

techniques are also combined, though in a different manner, to fit a more complex model.

3.6 Random Effects Models and Empirical Bayes Estimation

3.6.1 Preliminaries

As indicated in section 3.3.3, random effects models are appropriate if one is interested in cluster-specific inferences. Such models can be applied in the longitudinal setting, but are especially appropriate for cross-sectionally clustered data. An early application of a random effects model to binary data is given by Korn and Whittemore (1979). More general references include the previously mentioned papers by Laird and Ware (1982) and Stiratelli et al. (1984); see also Zeger et al. (1988). A good overview of the use of random effects models in the class of generalized linear models is given in Chapter 9 of Diggle et al. (1994).

Suppose that as in section 3.5.1 our data consist of observations Y_{ij}, $j = 1, ..., n_i$ and $i = 1, ..., K$. Along with the $p \times 1$ vector of fixed effects covariates x_{ij}, associate with each Y_{ij} a vector of r random effects covariates, contained in z_{ij}. This implies that the coefficients for the x_{ij} are constant across clusters, whereas the coefficients for the z_{ij}, i.e. the random effects, display heterogeneity across clusters, thus inducing within-cluster correlation. Similar to the development in section 3.5.1, in the GLM framework the density of Y_{ij} conditional on some (possibly vector-valued) random variable b_i is assumed to follow the exponential family form

$$f(y_{ij}|b_i) = \exp\{(y_{ij}\theta_{ij} - a(\theta_{ij}) + b(y_{ij}))\phi\}. \tag{3.6.1}$$

Conditional on b_i the observations $Y_{i1}, ..., Y_{in_i}$ are assumed to be independent. As before

$\mu_{ij} = \mathrm{E}(Y_{ij}|\boldsymbol{b}_i) = a'(\theta_{ij})$ and $\mathrm{Var}(Y_{ij}|\boldsymbol{b}_i) = a''(\theta_{ij})/\phi$, where now

$$g(\mu_{ij}) = \eta_{ij}, \quad \eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i, \tag{3.6.2}$$

for some link function $g(.)$. We assume that $\boldsymbol{b}_1, ..., \boldsymbol{b}_K$ are i.i.d. observations from some distribution $G(\boldsymbol{b})$ with mean 0 and covariance matrix Σ. The joint distribution of y_{ij} and $\boldsymbol{b}_i$ is $f(y_{ij}|\boldsymbol{b}_i)g(\boldsymbol{b}_i)$, so the marginal likelihood for the data is given by

$$L(\boldsymbol{\beta}, \Sigma; \mathbf{y}) = \prod_{i=1}^{K} \int \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{b}_i)g(\boldsymbol{b}_i)d\boldsymbol{b}_i. \tag{3.6.3}$$

For most non-linear models this integral is intractable and must be evaluated numerically. A fairly large literature is devoted to the non-trivial problem of parameter estimation in the generalized linear mixed model. References include Conaway (1990), who proposes a special model which yields a marginal likelihood that can be evaluated without numerical integration, Wolfinger and O'Connell (1993) who suggest a pseudo-likelihood approach, Waclawiw and Liang (1993, 1994) whose approach was discussed above, and Gibbons and Hedeker (1994) who propose a general random effects probit model. An excellent overview, discussing both theoretical issues and applications, is given by Breslow and Clayton (1993).

3.6.2 The Empirical Bayes Approach

Empirical Bayes approaches to inference for random effects models avoid the integration involved in (3.6.3). Instead of attempting to obtain the marginal likelihood, the problem is viewed from a Bayesian perspective. A noninformative prior distribution is placed on the fixed effects parameters, together with a proper prior for the random effects, and the resulting posterior distribution of all parameters is maximized. This procedure implies that the parameter(s) of the prior random effects distribution are known. In practice one could either postulate a hyperprior for these parameters, yielding a fully Bayesian analysis,

or estimate them empirically in some fashion, giving rise to the empirical Bayes approach. The latter will be the focus for the remainder of this section, in preparation for the model discussed in Chapter 4. There is a large literature documenting the successful applications of this approach; samples include the work of Morris (1983), MacGibbon and Tomberlin (1989), Farrell (1991) and Farrell et al. (1994).

To fix ideas more firmly, note that we can write the joint distribution of the data, the fixed and the random effects as

$$\begin{aligned}
\prod_{i=1}^{K}\{\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{b}_i)g(\boldsymbol{b}_i)\} &\propto \prod_{i=1}^{K}\{\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{b}_i,\boldsymbol{\beta},\Sigma)g(\boldsymbol{b}_i,\boldsymbol{\beta}|\Sigma)\} \\
&= [\prod_{i=1}^{K}\{\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{b}_i,\boldsymbol{\beta},\Sigma)\}\,[\prod_{i=1}^{K} g(\boldsymbol{b}_i,\boldsymbol{\beta}|\Sigma)] \\
&= p(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\Sigma)p(\boldsymbol{b},\boldsymbol{\beta}|\Sigma) \\
&= p(\boldsymbol{y},\boldsymbol{b},\boldsymbol{\beta}|\Sigma) \qquad (3.6.4)
\end{aligned}$$

The right side of the first line in (3.6.4) is just a re-expression of the left side which emphasizes the assumed dependencies in the conditional and prior distributions. Observe that we can write $g(\boldsymbol{b}_i) \propto g(\boldsymbol{b}_i,\boldsymbol{\beta})$ since a noninformative prior distribution is assumed for $\boldsymbol{\beta}$. The posterior distribution of the parameters given the data is then

$$p(\boldsymbol{b},\boldsymbol{\beta}|\boldsymbol{y},\Sigma) = \frac{p(\boldsymbol{y},\boldsymbol{b},\boldsymbol{\beta}|\Sigma)}{p(\boldsymbol{y}|\Sigma)}. \qquad (3.6.5)$$

In most cases finding a closed-form expression for equation (3.6.5) is an intractable problem. But we can nevertheless maximize the posterior in $(\boldsymbol{b},\boldsymbol{\beta})$ (or equivalently the log-posterior) by simply maximizing the joint distribution, since the denominator in (3.6.5) does not involve these parameters. As is standard practice, and stated as follows for instance in MacGibbon and Tomberlin (1989), the posterior is expressed as a multivariate normal distribution having its mean at the mode of (3.6.5) and variance equal to the inverse of

the information matrix evaluated at the mode.

Now all of the distributions in the above equations assumed that the prior covariance matrix Σ is known. Since this is not the case in practice, the empirical Bayes approach uses an EM-type algorithm (Dempster, Laird and Rubin (1977)) to estimate Σ. Beginning with an initial estimate $\Sigma^{(0)}$, (3.6.5) or equivalently the joint distribution (3.6.4) is maximized, yielding $(\hat{b}, \hat{\beta})$ and $\hat{\text{Cov}}((\hat{b}, \hat{\beta}))$. From these estimates an updated value for Σ is computed, following the same argument as outlined for the univariate case, referring back to equation (3.5.26). Assuming a zero-mean prior distribution,

$$\begin{aligned}
\Sigma = \text{Cov}(b_i) &= \text{E}(b_i b_i^T) \\
&= \text{E}((\hat{b}_i - (\hat{b}_i - b_i))(\hat{b}_i - (\hat{b}_i - b_i))^T) \\
&= \text{E}(\hat{b}_i \hat{b}_i^T - \hat{b}_i(\hat{b}_i - b_i)^T - (\hat{b}_i - b_i)\hat{b}_i^T + (\hat{b}_i - b_i)(\hat{b}_i - b_i)^T) \\
&\simeq \text{E}(\hat{b}_i \hat{b}_i^T) + \text{E}(\hat{b}_i - b_i)(\hat{b}_i - b_i)^T. \qquad (3.6.6)
\end{aligned}$$

Hence (3.6.6) can be estimated by the sum of $K^{-1}\sum_{i=1}^{K} \hat{b}_i \hat{b}_i^T$ and the estimated average asymptotic posterior covariance matrix for $\hat{b}_i$, $i = 1, ..., K$. The updated value for Σ is therefore

$$\Sigma^{(1)} = \frac{\sum_{i=1}^{K} \hat{b}_i \hat{b}_i^T}{K} + \frac{\sum_{i=1}^{K} \hat{\text{Cov}}(\hat{b}_i | y)}{K}. \qquad (3.6.7)$$

Given $\Sigma^{(1)}$ new parameter estimates are obtained from the maximization of (3.6.4), whereupon the estimate of Σ is updated again, and so forth, until convergence is achieved.

3.6.3 Some Examples

We now describe the standard logistic-normal random effects model for the case of a univariate random effects distribution, and will discuss two examples where an extension to

either two separate or a bivariate distribution is appropriate. In the next chapter we will introduce a generalization of this standard model to allow for the simultaneous modelling of cross-sectional and longitudinal dependence.

As already outlined in section 3.5.6, the logistic-normal model for cluster correlated binary data, assuming a single (scalar) random effect for each cluster, can be stated as follows:

$$\begin{aligned} Y_{ij}|x_{ij}, b_i &\sim \text{Bin}(1, p_{ij}), \qquad \text{where} \\ \theta_{ij} &= \log(\frac{p_{ij}}{1-p_{ij}}) = x_{ij}'\beta + b_i, \\ b_i &\overset{\text{iid}}{\sim} N(0, \sigma^2). \end{aligned} \tag{3.6.8}$$

The log of the posterior distribution is given by

$$\log p(b, \beta|y, \sigma^2) \propto \sum_{i=1}^{K}\sum_{j=1}^{n_i}\{y_{ij}(x_{ij}'\beta + b_i) - \log(1 + e^{x_{ij}'\beta + b_i})\} - \frac{1}{2\sigma^2}\sum_{i=1}^{K} b_i^2. \tag{3.6.9}$$

This function is easily maximized, given a value of σ^2; see section 4.2.1 for details on estimation. The above model would be suitable for instance for a cross-sectional analysis of the WSPP3 cohort at one given time point, with clusters defined by schools.

Extensions of model (3.6.8) are relatively straightforward. Consider for example the survival model described in section 3.5.3. It was suggested there to use GEE to account for the clustering due to schools. If the analysis were to involve both elementary and highschools it would be sensible to estimate two exchangeable correlation parameters, one for each type of school, in the manner discussed in section 3.5.2. This would allow for a varying degree of similarity in the smoking behaviour of students from the same elemtary school as compared to students from the same highschool. Alternatively, an analogous

random effects model can be specified by postulating two independent random effects distributions, one for each type of school. Both would be centered at zero but would have unequal variances. (Indicator variables in the fixed effects portion of the model can be used to adjust for the difference in mean response rates over time). To write this model down we need a slight generalization of the notation in section 3.5.3. The definition of $Y_{ij}(r)$ in equation (3.5.14) implicitly assumed that the configuration of clusters is fixed across time, i.e. that individuals do not move to a different cluster (school) over the period of observation. In fact however, the school an individual is attending at any given time does depend on time, the most obvious example of this being the fact that students move from elementary to highschools after grade 8. Therefore, letting $t = 1, 2, 3, ...$ be the times of data collection, simply define T_i as the time to first smoking of the ith individual in the data set, and Y_{it} as

$$Y_{it} = \begin{cases} 1 & \text{if } T_i \leq t, \text{ given } T_i > t-1 \\ 0 & \text{if } T_i > t. \end{cases}$$

Further, let $b_{(it)}$ be the random effect associated with the school attended by individual i at time t, and let $E = \{b_{E1}, ..., b_{Ek_1}\}$ and $H = \{b_{H1}, ..., b_{Hk_2}\}$ be the sets of elementary and highschool random effects, respectively. Then the model can be stated as

$$\begin{aligned} Y_{it}|x_{it}, b_{(it)} &\sim \text{Bin}(1, p_{it}) \quad \text{where} \\ \log(\frac{p_{it}}{1-p_{it}}) &= x'_{it}\beta + b_{(it)}, \\ b_{(it)} &\sim \begin{cases} N(0, \sigma_E^2) & \text{if } b_{(it)} \in E \\ N(0, \sigma_H^2) & \text{if } b_{(it)} \in H. \end{cases} \end{aligned} \tag{3.6.10}$$

Assuming we have data on N individuals, each contributing $m_i \geq 1$ observations, the

log-posterior is therefore

$$\log p(b,\beta|y,\sigma_E^2,\sigma_H^2) \propto \sum_{i=1}^{N}\sum_{r=1}^{m_i}\{y_{ir}(x_{ir}'\beta + b_{(ir)}) - \log(1+e^{x_{ir}'\beta+b_{(ir)}})\} - \frac{1}{2\sigma_E^2}\sum_{k\in E} b_{Ek}^2 - \frac{1}{2\sigma_H^2}\sum_{k\in H} b_{Hk}^2. \quad (3.6.11)$$

Note that as above (3.6.11) assumes a logistic model for the conditional distribution of the data given the random effects. The complementary log-log link could be substituted just as easily if the proportional hazards model is of interest. The log-posterior would then be given by

$$\log p(b,\beta|y,\sigma_E^2,\sigma_H^2) \propto \sum_{i=1}^{N}\sum_{r=1}^{m_i}\{y_{ir}\log(1-\exp(-e^{x_{ir}'\beta+b_{(ir)}})) - (1-y_{ir})e^{x_{ir}'\beta+b_{(ir)}}\} - \frac{1}{2\sigma_E^2}\sum_{k\in E} b_{Ek}^2 - \frac{1}{2\sigma_H^2}\sum_{k\in H} b_{Hk}^2. \quad (3.6.12)$$

The univariate analogue of equation (3.6.7) is used to update separately the estimates of σ_E^2 and σ_H^2; observe that averages are taken over the k_1 elementary schools only to compute $\hat{\sigma^2}_E$, and over the k_2 highschools to compute $\hat{\sigma^2}_H$.

The second example we consider involves estimating a bivariate random effect for each school, allowing for correlation between the two marginal components. This would be an appropriate analogous model for example to the generalization described in section 3.5.2, concerning the estimation of separate correlation parameters for pairs of observations from the same gender group within a school, as well as for mixed gender pairs. For the sake of clarity consider data from a single time point, so that Y_{ij} will once again refer to individual j in school i. Let $z_{ij} = (1,0)^T$ if Y_{ij} is an observation on a female student, and $(0,1)^T$ if it is an observation on a male student. In addition, let $b_i = (b_{i1}, b_{i2})^T$, where b_{i1} and b_{i2} are the random effects associated with females and males, respectively, in school i. Suppressing

the dependence of Y_{ij} on the z_{ij} and the usual fixed effects covariate vector x_{ij}, the model is given by

$$\begin{aligned} Y_{ij}|\, b_i &\sim \text{Bin}(1, p_{ij}) \quad \text{where} \\ \log(\frac{p_{ij}}{1-p_{ij}}) &= x_{ij}'\beta + z_{ij}'b_i, \\ b_i &\overset{\text{iid}}{\sim} BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_F^2 & \rho\sigma_F\sigma_M \\ \rho\sigma_F\sigma_M & \sigma_M^2 \end{bmatrix}\right). \end{aligned} \tag{3.6.13}$$

Equality between the marginal variances σ_F^2 and σ_M^2 would suggest that the impact of school environment on the smoking behaviour of males and females is similar. One might expect such a finding to be accompanied by a strong positive correlation ρ, which indicates the extent to which males and females behave similarly within a school. Thus for example, it would be unusual to encounter numerous schools in which a large proportion of the boys were smoking, but not many of the girls, and vice versa. The log-posterior for model (3.6.13) is

$$\begin{aligned} \log p(b, \beta|y, \sigma_F^2, \sigma_M^2, \rho) \;\propto\; & \sum_{i=1}^{K}\sum_{j=1}^{n_i}\{y_{ij}(x_{ij}'\beta + z_{ij}'b_i) - \log(1 + e^{x_{ij}'\beta + z_{ij}'b_i})\} \\ & -\frac{1}{2(1-\rho^2)}\sum_{i=1}^{K}\{\frac{b_{i1}^2}{\sigma_F^2} - 2\rho\frac{b_{i1}b_{i2}}{\sigma_F\sigma_M} + \frac{b_{i2}^2}{\sigma_M^2}\}. \end{aligned} \tag{3.6.14}$$

Updated estimates of σ_F^2, σ_M^2 and ρ are obtained from the appropriate elements of updated covariance matrix, computed directly using (3.6.7).

With reference to these examples, in Chapter 5 we describe an alternative general approach to extending model (3.6.8).

A limitation of all of the approaches outlined so far is the fact that data exhibiting two levels of dependence, i.e. longitudinal and cross-sectional, cannot be analyzed in full

generality. In the next chapter we discuss this problem further and describe a method for simultaneously modelling data with such a composite correlation structure.

Chapter 4

Simultaneous Modelling of Cross-Sectional and Longitudinal Dependence

4.1 Introduction

The discussion thus far has pointed out several times that some studies give rise to data which can be logically grouped into clusters in more than one way. We have indicated that the WSPP3 study is a good example of a longitudinal study where the data are also cross-sectionally grouped in meaningful units. Clusters are thus defined for groups of observations belonging to the same unit (which may include data on the same individual at more than one time point) as well as for groups of observations made on the same individual over time.

If interest lies in a marginal analysis focussing only on one time point, one might analyze the data using GEE, specifying, for example, an exchangeable correlation structure to model the correlation expected to exist between students attending the same school.

Alternatively, the standard logistic-normal random effects model would achieve a similar end from the point of view of a conditional analysis. However if one wishes to analyze data from two or more observation times simultaneously, the longitudinal correlation must also be taken into account. One approach would be to carry on with GEE and merely specify a more complicated correlation structure: one might for example assume strong dependence between observations on the same individual, moderate correlation between observations on different individuals in the same school at the same time and still weaker correlation between observations on different individuals in the same school at different times. There are however two problems with this approach. Firstly, solving the estimating equations would require the repeated definition and inversion of potentially very large matrices, since cluster sizes in this framework are determined by the number of students in a given school times the number of observation points. The second and more limiting problem is the implicit assumption that all students within a given school stay in that school over all observation times. This approach does not allow the possibility of students changing schools over time, which is, however, a characteristic observed in the WSPP3 data. (Note in fact that apart from haphazard switching, students all change schools by design after grade 8, as they move from elementary to highschools).

4.2 A Composite Model

We present a composite analysis which avoids the problems outlined above. We consider the GEEs one would use for a straightforward longitudinal analysis and augment the linear predictor associated with each observation with a random effect for the school attended by that particular individual at the particular time in question. It is assumed that the effect of school environment is common to all observations collected in a given school, i.e. not just those gathered at one particular time point. Observations from different individuals in the same school are taken to be independent conditional on the random effect for that school.

Note that there is no restriction on which school an individual attends at any given time. The school random effects will be estimated using empirical Bayes methods, as described in section 3.6.

In the following two sections we first review estimation for the logistic-normal random effects model, assuming only one level of clustering in the cross-sectional sense, and then describe the more general composite model, which we refer to as the 'quasi empirical Bayes model'.

4.2.1 Estimation in the Standard Logistic-Normal Empirical Bayes Random Effects Model

Consider model (3.6.8) for data collected in K schools, with n_i observations from school i. The empirical Bayes estimating equations for this model can be expressed as

$$\begin{aligned} U_{EB}(\beta) &= \sum_{i=1}^{K} X_i^T (Y_i - p_i) = 0, \\ U_{EB}(b_i) &= 1^T (Y_i - p_i) - \frac{b_i}{\sigma^2} = 0, \\ & \qquad i = 1, ..., K \end{aligned} \tag{4.2.1}$$

where X_i is the design matrix for the ith cluster (school), Y_i and p_i are $(Y_{i1}, ..., Y_{in_i})^T$ and $(p_{i1}, ..., p_{in_i})^T$ respectively, and 1 is a unit vector of appropriate dimension. The matrix of negative second derivatives of the log posterior, whose inverse serves as an approximate covariance matrix in much the same way as the information matrix in a likelihood analysis, is of the form

$$I_{EB} = \begin{bmatrix} A_{p \times p} & B_{p \times K} \\ B^T_{K \times p} & C_{K \times K} \end{bmatrix} \tag{4.2.2}$$

with

$$\begin{aligned} A_{p\times p} &= \sum_{i=1}^{K} X_i^T A_i X_i \\ B_{p\times K} &= (B_1, ..., B_K), \\ & \quad B_i = X_i^T A_i \, 1 \\ C_{K\times K} &= \text{diag}\{C_1, ..., C_K\}, \\ & \quad C_i = 1^T A_i \, 1 + \frac{1}{\sigma^2}, \end{aligned}$$

where $A_i = \text{diag}\{p_{i1}(1-p_{i1}), ..., p_{in_i}(1-p_{in_i})\}$. For a given value of the prior variance σ^2 the system of equations (4.2.1) is solved to produce estimates of the fixed and random effects, using the standard Newton-Raphson algorithm

$$\hat{\gamma}^{(\ell+1)} = \hat{\gamma}^{(\ell)} + I_{EB}^{-1} \, U_{EB}|_{\gamma=\hat{\gamma}^{(\ell)}},$$

where $\hat{\gamma}^{(\ell)}$ is the estimate of $\gamma = (\beta, b_1, ..., b_K)^T$ after the ℓth iteration, and $U_{EB} = (U_{EB}(\beta), U_{EB}(b_1), ..., U_{EB}(b_K))^T$.

The univariate version of (3.6.7) is used to update the estimate of the prior variance, and with this estimate $U_{EB} = 0$ is solved once again to produce new fixed and random effects estimates. This cycle continues until convergence in σ^2 is achieved.

4.2.2 The Quasi Empirical Bayes Model

Liang and Zeger (1986) moved from likelihood score equations to generalized estimating equations for longitudinal data ((3.5.3) to (3.5.7)) by introducing a working covariance matrix, reflecting the correlation structure among repeated observations on the same individual, into the estimating equations. We extend the empirical Bayes estimating equations U_{EB} in a similar manner in order to facilitate the analysis of data exhibiting both a cross-

sectional as well as a longitudinal component of clustering.

The notation used thus far has been largely dependent on the specific context under consideration, and in what follows it will be crucial to be able to clearly distinguish observations or groups of observations as belonging either to an individual or to a school. Unless otherwise indicated, the following notation will therefore be strictly adhered to throughout this chapter. Let the data consist of observations on N individuals (students), each observed at T times. Let i be the subscript used to refer to an individual, and t the subscript used to refer to a time point. Furthermore let the $N \times T$ observations be collected in K schools, and let the subscript k refer to a school, with j indicating the jth observation within a school. Thus Y_{it} denotes the tth observation on individual i ($t = 1, ..., T$, $i = 1, ..., N$), whereas Y_{kj} denotes the jth observation in school k ($j = 1, ..., n_k$, $k = 1, ..., K$). Similarly, vectors or matrices with a single subscript i or k refer to collections of observations on the corresponding individual or school, respectively.

Consider the following model:

$$Y_{it}|x_{it}, b_{(it)} \sim \text{Bin}(1, p_{it}), \qquad \text{Corr}(Y_{it}, Y_{it'}|\{b_{(it)}, b_{(it')}\}) = \rho_{tt'},$$

where

$$\begin{aligned} \log(\frac{p_{it}}{1-p_{it}}) &= x'_{it}\beta + b_{(it)}, \\ b_{(it)} &\in \{b_1, ..., b_K\}, \\ b_k &\sim N(0, \sigma^2), \quad k = 1, ..., K. \end{aligned} \tag{4.2.3}$$

Here $b_{(it)}$ is the random effect associated with the school attended by individual i at time t. The model assumes that the correlation between two observations on the same individual is only a function of time, conditional on the effect of the school(s) attended by that individual at the two time points.

We are interested in estimating the fixed effect parameters β and within-individual correlations $\rho = (\rho_{12}, ..., \rho_{T-1T})$, as well as the random effects b_k, $k = 1, ..., K$ and their variance σ^2. To accomplish this we generalize the system of equations in (4.2.1) as follows: note that $U_{EB}(\beta)$ can be written either as a sum over schools or as a sum over individuals. (In fact, it just equals $X^T(Y - p)$ where X is the full design matrix, and Y is the corresponding response vector, with $p = \mathrm{E}(Y)$). To incorporate the correlation among repeated observations on the same individual into the estimating procedure, construct the estimating equations for the fixed effects as sums over individuals, introducing a working covariance matrix in the same manner as described in section 3.5.1. As indicated at the beginning of section 4.2 this yields GEEs of exactly the same form as (3.5.7), with the exception that the probability p_{it} is conditional on $b_{(it)}$. We therefore write

$$U_{QEB}(\beta) = \sum_{i=1}^{N} X_i^T A_i V_i^{-1} (Y_i - p_i), \tag{4.2.4}$$

where $A_i = \mathrm{diag}\{p_{i1}(1 - p_{i1}), ..., p_{iT}(1 - p_{iT})\}$ and V_i is the working covariance matrix $A_i^{1/2} R(\alpha) A_i^{1/2}$. As usual α represents all parameters required to specify the working correlation matrix R; we can write $\alpha = \rho$. The estimating equations for the random effects have a similar form; they can be written as

$$U_{QEB}(b_k) = 1^T \mathcal{A}_k \mathcal{V}_k^{-1} (Y_k - p_k) - \frac{b_k}{\sigma^2}, \quad k = 1, ..., K \tag{4.2.5}$$

where $\mathcal{A}_k = \mathrm{diag}\{p_{k1}(1 - p_{k1}), ..., p_{kn_k}(1 - p_{kn_k})\}$ and $\mathcal{V}_k$ is the working covariance matrix $\mathcal{A}_k^{1/2} \mathcal{R}_k(\alpha) \mathcal{A}_k^{1/2}$ defined for school k. The entries of the corresponding correlation matrix $\mathcal{R}_k(\alpha)$ are defined in terms of the conditional correlations between repeated observations

on the same individual in the same school, given the random effect for that school:

$$\text{Corr}(Y_{kj}, Y_{kj'}|b_k) = [\mathcal{R}_k(\alpha)]_{jj'} = \begin{cases} \rho_{tt'} & \text{if } Y_{kj} \text{ and } Y_{kj'} \text{ are obs'ns taken on the} \\ & \text{same student, at times } t \text{ and } t' \\ 0 & \text{otherwise} \end{cases}$$
$$k = 1, ..., K \qquad (4.2.6)$$

The matrix of negative second derivatives is given by

$$I_{QBB} = \begin{bmatrix} A_{p\times p} & B_{p\times K} \\ B^T_{K\times p} & C_{K\times K} \end{bmatrix} \qquad (4.2.7)$$

where

$$\begin{aligned} A_{p\times p} &= \sum_{i=1}^{N} X_i^T A_i V_i^{-1} A_i X_i \\ B_{p\times K} &= (B_1, ..., B_K), \\ & \quad B_k = X_k^T A_k \mathcal{V}_k^{-1} A_k \, 1 \\ C_{K\times K} &= \text{diag}\{C_1, ..., C_K\}, \\ & \quad C_k = 1^T A_k \mathcal{V}_k^{-1} A_k \, 1 + \frac{1}{\sigma^2}. \end{aligned}$$

For a given value of σ^2 the quasi empirical Bayes estimating equations

$$U_{QBB} = (U_{QBB}(\beta), U_{QBB}(b_1), ..., U_{QBB}(b_K))^T$$

are set to zero to solve for the estimates of the fixed and random effects. Moment estimates for the intra-individual correlation parameters ρ are computed after each iteration toward a solution to the estimating equations (see section 3.5 or Liang and Zeger (1986)). If the number of repeated observations on each individual is relatively small, we suggest using an

unspecified correlation structure for $R(\alpha)$. Hence we define this $T \times T$ working correlation matrix as having diagonal elements equal to 1, and off-diagonal elements $\rho_{tt'}$ estimated by the off-diagonal elements of the matrix

$$(N-p)^{-1} \sum_{i=1}^{N} A_i^{-1/2}(Y_i - p_i)(Y_i - p_i)^T A_i^{-1/2}.$$

After the nth estimation cycle conditional on $\sigma^2_{(n)}$, the prior variance is updated using the formula

$$\sigma^2_{(n+1)} = \frac{\sum_{k=1}^{K} b^2_{k(n)}}{K} + \frac{\sum_{k=1}^{K} \hat{\text{Var}}(b_{k(n)}|y)}{K}, \tag{4.2.8}$$

where $\widehat{\text{Var}}(b_{k(n)}|y)$ is the $(p+k, p+k)$ element of I^{-1}_{QEB}, evaluated at the nth cycle.

4.3 A Robust Covariance Matrix

4.3.1 The Problem

Model-based variance estimates for the estimates $(\hat{\beta}, \hat{b}_1, ..., \hat{b}_K)$ from the quasi empirical Bayes model are obtained from the diagonal elements of $\hat{I}^{-1}_{QEB}$. Observe however that these estimates are computed assuming that the value of the random effects variance, on which they are conditioned, is fixed. They do not account for the additional variability induced by the fact that σ^2 is empirically estimated. Bootstrap adjustments to the model-based variance estimates are commonly used with empirical Bayes procedures to correct this problem, yielding robust variance estimates. Refer for example to the detailed exposition given in Laird and Louis (1987), or to Farrell (1991); Waclawiw and Liang (1994) also combine parametric bootstrap procedures with their estimating function approach.

The bootstrap is much harder to apply in a similar manner when the data are clustered in more than one direction. The parametric boostrap would require generating data

exhibiting the same underlying correlation structure as seen in the original data, which, except in simple circumstances, is virtually impossible. On the other hand, a nonparametric bootstrap procedure is not feasible either, since the fact that individual students can change schools over time implies that observations belonging to the same individual could appear in more than one school, thus precluding schools as the resampling unit.

We adopt a different approach to generate a robust covariance matrix, which avoids resampling altogether. Suppose that instead of empirical Bayes estimation we had carried out a fully Bayesian analysis which assumes the postulation of a hyperprior distribution for σ^2, whose parameters are known or in turn somehow estimated. The system of equations we would have solved to obtain our parameter estimates would have included equations for both fixed and random effects, as well as the prior variance σ^2. The inverse of the matrix of negative second derivatives derived from the fully Bayesian approach would yield variance estimates duly adjusted for the estimation of all parameters in the model (except the hyperprior parameters, which need not concern us here), including σ^2. Now let us define a Bayesian analysis to be *equivalent* to an empirical Bayes procedure for the same problem if the two analyses yield the same estimates for the fixed and random effects, as well as for the random effects variance. It seems reasonable therefore to determine, for the results of a given empirical Bayes analysis, an equivalent Bayesian analysis in the above sense, and use the inverse matrix of negative second derivatives from this as a robust covariance matrix. An added advantage of proceeding in this manner is that an estimate of the variability in $\hat{\sigma^2}$ is obtained in the process, which is not available from the empirical Bayes estimation alone.

4.3.2 A Bayesian Formulation Equivalent to Empirical Bayes

Assume for the moment that repeated observations on the same individual are conditionally independent, so that the data are cross-sectionally clustered only, putting us back in the standard empiricial Bayes framework. We make this distinction for the sake of clarity and to be able to write down appropriate posterior distributions. The following development applies in a similar manner to the quasi empirical Bayes model, and extends immediately to it.

Consider now a fully Bayesian analysis of the logistic-normal model, which may be expressed as follows:

$$\begin{aligned} Y_{kj}|x_{kj}, b_k &\sim \text{Bin}(1, p_{kj}) \quad \text{where} \\ \log(\frac{p_{kj}}{1-p_{kj}}) &= x'_{kj}\beta + b_k, \\ b_k &\overset{\text{iid}}{\sim} N(0, \sigma^2), \\ \sigma^2 &\sim G(\sigma^2, \nu). \end{aligned} \tag{4.3.1}$$

Let $G(\sigma^2, \nu)$ represent the C.D.F. of the hyperprior distribution on the random effects variance, and denote its density by $g(\sigma^2, \nu)$. We assume this distribution is fully specified by the (possibly vector-valued) hyperparameter ν. The posterior distribution for the fixed and random effects, as well as their variance, is given by

$$\begin{aligned} &p(\beta, b, \sigma^2|y, \nu) \propto \\ &\quad \{\prod_{k=1}^{K}\prod_{j=1}^{n_k} p_{kj}^{y_{kj}}(1-p_{kj})^{1-y_{kj}}\}(\frac{1}{\sqrt{2\pi}\sigma})^K \exp\{-\frac{1}{2\sigma^2}\sum_{k=1}^{K} b_k^2\} g(\sigma^2, \nu). \end{aligned} \tag{4.3.2}$$

We are interested in a distribution $G(\sigma^2, \nu)$ which will yield estimates equivalent to those obtained from the analogous empirical Bayes formulation. Although the choice of this dis-

tribution is certainly not unique in the class of distributions with support on the positive real line, it is also not arbitrary. The following proposition illustrates this with reference to the exponential distribution.

Propositon 4.3.1: The exponential distribution cannot serve as a hyperprior for σ^2 in a fully Bayesian formulation of the logistic-normal model, if this analysis is to be equivalent to an empirical Bayes analysis.

Proof: Assume that there exists an exponential distribution which yields the same estimates when used as a hyperprior in a fully Bayesian analysis, as does the corresponding empirical Bayes analysis. Then

$$g(\sigma^2, \nu) = \nu e^{-\nu\sigma^2}$$

for some $\nu > 0$. The estimating function used to obtain the estimate of σ^2 is

$$\frac{\partial \log p(\boldsymbol{\beta}, \boldsymbol{b}, \sigma^2 | \boldsymbol{y}, \nu)}{\partial \sigma^2} = -\frac{K}{2\sigma^2} + \frac{\sum_{k=1}^{K} b_k^2}{2\sigma^4} - \nu. \tag{4.3.3}$$

Let $\hat{b}_{kB}$ and $\hat{\sigma^2}_B$ be the Bayesian estimates of b_k and σ^2, and let $\hat{b}_{kEB}$ and $\hat{\sigma^2}_{EB}$ be the empirical Bayes estimates. Evaluated at $b_k = \hat{b}_{kB}$ and $\sigma^2 = \hat{\sigma^2}_B$, equation (4.3.3) equals zero, yielding the following quadratic equation in $\hat{\sigma^2}_B$:

$$\hat{\sigma^2}_B = \frac{1}{K}\left(\sum_{k=1}^{K} \hat{b}_{kB}^2 - 2\nu\hat{\sigma^4}_B\right). \tag{4.3.4}$$

Now if the Bayesian and the empirical Bayes analyses were equivalent we would have

$$\hat{b}_{kB} = \hat{b}_{kEB} \quad \text{and} \quad \hat{\sigma^2}_B = \hat{\sigma^2}_{EB}, \quad k = 1, ..., K.$$

However

$$\begin{aligned}\hat{\sigma^2}_{EB} &= \frac{1}{K}(\sum_{k=1}^{K}\hat{b}^2_{kEB} + \sum_{k=1}^{K}\hat{\mathrm{Var}}(\hat{b}_{kEB}|y)) \\ &= \frac{1}{K}(\sum_{k=1}^{K}\hat{b}^2_{kEB} + m_1),\end{aligned}$$

where $m_1 > 0$. Furthermore from (4.3.4) we observe that

$$\hat{\sigma^2}_B = \frac{1}{K}(\sum_{k=1}^{K}\hat{b}^2_{kB} + m_2),$$

where $m_2 < 0$ since $\nu > 0$. Therefore if the random effects estimates from the two analyses were equivalent, this would imply that $\hat{\sigma^2}_{EB} \neq \hat{\sigma^2}_B$; similarly if the two estimates of σ^2 were equivalent, this would imply that the random effects estimates could not be. Hence an exponential hyperprior cannot be used in a Bayesian analysis, if this is to be interpreted as equivalent to an empirical Bayes analysis.

A similar result holds for the Pareto distribution, and we conjecture that it holds for any distribution on positive support with mode strictly at zero.

We therefore require a more general hyperprior. In our experience the gamma distribution is a suitable choice, though other families exhibiting similar shapes, such as the lognormal, generalized Pareto or Weibull, may be equally useful. Even within the class of gamma distributions there are undoubtedly various parameter choices which would yield the same result. Here we shall focus specifically on the gamma$(\frac{\nu}{2}, \frac{1}{2})$ distribution with density

$$g(\sigma^2, \nu) = \frac{(\sigma^2)^{\nu/2-1}\exp\{-\sigma^2/2\}}{2^{\nu/2}\Gamma(\frac{\nu}{2})}, \quad \nu > 0. \tag{4.3.5}$$

Note that for positive integers ν, (4.3.5) is the density of a chi-square distribution on ν

degrees of freedom. In any case the mean and variance of $G \sim \text{gamma}(\frac{\nu}{2}, \frac{1}{2})$ are ν and 2ν respectively. The following proposition states the central result of this section.

Propositon 4.3.2: The empirical Bayes analysis for the logistic-normal model is equivalent to a fully Bayesian analysis assuming a gamma$(\frac{\nu}{2}, \frac{1}{2})$ hyperprior, where

$$\nu = 2 + K + \hat{\sigma^2}_{EB} - \frac{\sum_{k=1}^{K} \hat{b}^2_{kEB}}{\hat{\sigma^2}_{EB}} \tag{4.3.6}$$

and $\hat{b}_{kEB}$ and $\hat{\sigma^2}_{EB}$ are the empirical Bayes estimates of b_k and the prior variance, respectively.

Proof: Substituting (4.3.5) into (4.3.2) we derive the following estimating function for σ^2:

$$U(\sigma^2) = \frac{\partial \log p(\beta, b, \sigma^2 | y, \nu)}{\partial \sigma^2} = -\frac{K}{2\sigma^2} + \frac{\sum_{k=1}^{K} b_k^2}{2\sigma^4} + \frac{\nu/2 - 1}{\sigma^2} - \frac{1}{2}. \tag{4.3.7}$$

Note that for a given value of σ^2 the estimating equations for the fixed and random effects from the fully Bayesian posterior are equivalent to those from the empirical Bayes posterior; hence they lead to the same estimates of these quantities. If in addition that value of σ^2 equals $\hat{\sigma^2}_{EB}$ then

$$\begin{aligned} U(\hat{\sigma^2}_{EB})|_{b_k = \hat{b}_{kEB}} &= -\frac{K}{2\hat{\sigma^2}_{EB}} + \frac{\sum_{k=1}^{K} \hat{b}^2_{kEB}}{2\hat{\sigma^4}_{EB}} + \frac{\nu - 2}{2\hat{\sigma^2}_{EB}} - \frac{1}{2} \\ &= \frac{-K\hat{\sigma^2}_{EB} + \sum_{k=1}^{K} \hat{b}^2_{kEB} + \hat{\sigma^4}_{EB} + K\hat{\sigma^2}_{EB} - \sum_{k=1}^{K} \hat{b}^2_{kEB} - \hat{\sigma^4}_{EB}}{2\hat{\sigma^4}_{EB}} \\ &= 0, \end{aligned}$$

implying that the empirical Bayes estimate of the prior variance is also a solution to the

Bayesian estimating equation for σ^2 and hence equal to $\hat{\sigma^2}_B$. The two analyses are therefore equivalent.

Since we have now found a hyperprior which will generate the same (Bayesian) estimates as an empirical Bayes analysis would, we can use the estimates from the empirical Bayes model to evaluate the fully Bayesian matrix of negative second derivatives, and use the diagonal elements of its inverse as variance estimates, adjusted for the estimation of σ^2. This matrix is simply I_{EB} (see (4.2.2)), augmented by another row and column corresponding to the negative second derivatives of the log-posterior with respect to σ^2 and β, σ^2 and b_k, and $(\sigma^2)^2$. It can be written as

$$I_B = \left[\begin{array}{c|c} I_{EB} & E \\ \hline E^T & D \end{array}\right], \tag{4.3.8}$$

where (letting $p_B = p(\beta, b, \sigma^2|y, \nu)$)

$$\begin{aligned} E^T &= (-\frac{\partial^2 \log p_B}{\partial\beta\partial\sigma^2}, -\frac{\partial^2 \log p_B}{\partial b_1 \partial\sigma^2}, ..., -\frac{\partial^2 \log p_B}{\partial b_K \partial\sigma^2}) \\ &= (-\frac{\partial U(\sigma^2)}{\partial\beta}, -\frac{\partial U(\sigma^2)}{\partial b_1}, ..., -\frac{\partial U(\sigma^2)}{\partial b_K}) \\ &= (0, ..., 0, -\frac{b_1}{\sigma^4}, ..., -\frac{b_K}{\sigma^4})_{1\times(p+K)} \end{aligned}$$

and

$$\begin{aligned} D &= -\frac{\partial^2 \log p_B}{\partial(\sigma^2)^2} = -\frac{\partial U(\sigma^2)}{\partial\sigma^2} \\ &= \frac{\sum_{k=1}^K b_k^2}{\sigma^6} - \frac{K-\nu+2}{2\sigma^4}. \end{aligned}$$

As mentioned at the end of section 4.3.1, proceeding in this manner allows us to obtain a variance estimate for $\hat{\sigma^2}_{EB}$ as well. This is given by the $(p+K+1),(p+K+1)$ entry of I_B^{-1}. It can be explicitly expressed as

$$\widehat{\mathrm{Var}}(\hat{\sigma^2}_{EB}) = (\hat{D} - \hat{E}^T \hat{I}_{EB}^{-1} \hat{E})^{-1}, \tag{4.3.9}$$

where a 'hat' denotes evaluation at the empirical Bayes estimates.

As indicated earlier, the results of this section apply to the quasi empirical Bayes model in the same way as described here. Simply replace I_B with

$$I_{QB} = \left[\begin{array}{c|c} I_{QEB} & E \\ \hline E^T & D \end{array}\right] \tag{4.3.10}$$

and equation (4.3.9) with

$$\widehat{\mathrm{Var}}(\hat{\sigma^2}_{QEB}) = (\hat{D} - \hat{E}^T \hat{I}_{QEB}^{-1} \hat{E})^{-1}, \tag{4.3.11}$$

where a 'hat' now indicates evaluation at the quasi empirical Bayes estimates. Note in particular that the entries of E and D are unaffected by incorporating intra-individual correlations into the estimation procedure.

The performance of the robust variance estimates described here, including the variance estimate for $\hat{\sigma^2}$, will be studied in section 4.5.

4.4 Generating Data With a Specified Composite Correlation Structure

In this section we describe the simulation of various data scenarios, which will be used in the next section to study the properties of the estimates from the quasi empirical Bayes model. We shall consider three cases, for each of which we generate data sets consisting of three observations on each of 200 students, attending a total of 20 schools (10 students, or 30 observations, in each school). Note that for the purpose of the simulation study, students will in fact remain in the same school over all three observation times, so that $b_{(it)}$ in (4.2.3) remains constant as t varies. Hence in this case we can let Y_{kjt} represent the tth observation on individual j in school k. This is a slight alteration of the notation introduced in section 4.2.2 in that two subscripts are used to refer to an observation in a given school, denoting the individual responding and the time point in question. Model (4.2.3) can then be equivalently expressed as

$$\begin{aligned} Y_{kjt}|x_{kjt}, b_k &\sim \text{Bin}(1, p_{kjt}), \quad \text{where} \\ \text{Corr}(Y_{kjt}, Y_{kjt'}|b_k) &= \rho_{tt'}, \\ \log(\frac{p_{kjt}}{1-p_{kjt}}) &= x'_{kjt}\beta + b_k, \\ b_k &\sim N(0, \sigma^2), \quad k = 1, ..., K. \end{aligned} \tag{4.4.1}$$

It is only possible to express the model in this manner because of the present restriction that students do not change schools over time. For the rest of this section it will be more convenient and helpful to refer, not to the model as given in (4.2.3), but instead to its equivalent restatement in (4.4.1).

The first case we consider is the simplest and models the response Y_{kjt} as a function of

time only. Thus, referring to (4.4.1), we have

$$\log(\frac{p_{kjt}}{1-p_{kjt}}) = \beta_1 x_{kjt1} + \beta_2 x_{kjt2} + \beta_3 x_{kjt3} + b_k,$$

where

$$x_{kjtr} = \begin{cases} 1 & \text{if } t = r \\ 0 & \text{otherwise} \end{cases} \qquad r = 1, 2, 3.$$

The second case models Y_{kjt} as a function of time as well as a dichotomous individual-level covariate, assumed to be time-independent. We assume that half of the students in each school belong to one group and half to the other. This simulates a variable indicating gender, for example. Retaining x_{kjt1}, x_{kjt2} and x_{kjt3} as above, and defining g_{kj} as

$$g_{kj} = \begin{cases} 1 & \text{if student } j \text{ in school } k \text{ is female} \\ 0 & \text{otherwise} \end{cases}$$

we have in this case

$$\log(\frac{p_{kjt}}{1-p_{kjt}}) = \beta_1 x_{kjt1} + \beta_2 x_{kjt2} + \beta_3 x_{kjt3} + \beta_4 g_{kj} + b_k.$$

(To be fully consistent with (4.4.1) we could think of the fourth covariate g_{kj} as x_{kjt4}, constant across time.)

The third case models Y_{kjt} as a function of time as well as a dichotomous school-level covariate, also assumed to be time-independent. Here we assume that half of the schools under study belong to one group and half to the other. This simulates a variable indicating treatment condition, for example, where the intervention is applied at the level

of the school. Defining g_k as

$$g_k = \begin{cases} 1 & \text{if school } k \text{ belongs to the treatment group} \\ 0 & \text{if school } k \text{ belongs to the control group} \end{cases}$$

we have

$$\log(\frac{p_{kjt}}{1-p_{kjt}}) = \beta_1 x_{kjt1} + \beta_2 x_{kjt2} + \beta_3 x_{kjt3} + \beta_4 g_k + b_k.$$

(As above we could think of g_k as x_{kjt4}, now constant across time and subjects within schools.)

These data scenarios allow us to investigate the quasi empirical Bayes model for a number of common study designs. Thus the second case could be interpreted as a trial in which the data are clustered in schools but individuals are the units of randomization, being assigned to either treatment or control within schools. Observe that this case could also be interpreted as a matched pairs analysis. In each of 20 schools, we can think of half of the students as belonging to a treatment group, and half to a control group. We could equivalently view this as data on 40 schools, each half the size of the original 20, and divided on the treatment variable. Each of the 20 distinct random effects associated with the original school structure would now apply to exactly two of the new schools, creating 20 matched pairs of schools, with each pair containing one treatment and one control school. (This type of design allows treatment comparisons between schools which are largely unaffected by other known or unkown school-level covariates, due to the matching. It would make sense with such a study design to postulate random effects not for individual schools, but for the matched pairs of schools). Finally, as indicated above, the third case reflects an (unmatched) cluster-randomization trial in which schools are randomized to either a treatment or a control condition.

The actual data simulation is described in the Appendix at the end of this chapter. For each of the three cases discussed, the same 'true' parameter values were used to generate

the data. The fixed effects parameters were set to the following values:

$$\beta_1 = -2.0 \qquad \beta_2 = -1.0 \qquad \beta_3 = 0.0 \qquad \beta_4 = 1.0.$$

This implies an increasing probability of response ($Y = 1$) with time, and the same effect size associated with the individual-level covariate as with the school-level covariate. Since we are simulating data for three time points, we must specify three intra-individual correlation parameters:

$$\begin{aligned} \text{Corr}(Y_{kj1}, Y_{kj2}|b_k) &= \rho_{12} = 0.3 \\ \text{Corr}(Y_{kj2}, Y_{kj3}|b_k) &= \rho_{23} = 0.4 \\ \text{Corr}(Y_{kj1}, Y_{kj3}|b_k) &= \rho_{13} = 0.2. \end{aligned}$$

(Though not explicitly indicated, it is of course assumed that $\text{Corr}(Y_{kjt}, Y_{kjt'}|b_k)$ is conditional not only on the random effect for school k, but also on the covariate vectors x_{kjt} and $x_{kjt'}$). Emrich and Piedmonte (1991) point out that for correlated binary variables, $\rho_{tt'}$ is not free to vary over (-1,1), but must satisfy certain range restrictions to ensure that the joint probability function for the responses is nonnegative for all outcomes (see also Prentice (1988)). In particular, for given marginal probabilities p_t and $p_{t'}$, $\rho_{tt'}$ must satisfy

$$\rho_{tt'} \leq \min \left\{ \left(\frac{p_t(1-p_{t'})}{p_{t'}(1-p_t)} \right)^{1/2}, \left(\frac{p_{t'}(1-p_t)}{p_t(1-p_{t'})} \right)^{1/2} \right\}. \tag{4.4.2}$$

In other words, $\rho_{tt'}$ must be less than or equal to the root of the smaller of the two odds ratios that can be formed by p_t and $p_{t'}$, which is

$$\frac{p_t}{1-p_t} / \frac{p_{t'}}{1-p_{t'}} \quad \text{if } p_{t'} > p_t.$$

(A similar lower bound can be placed on $\rho_{tt'}$, but this bound is strictly negative and therefore not a concern in the present problem). It is easily verified that in all of the models we consider, the probabilities for any given individual at times 1 and 2 differ by 1 on the logistic scale, from which it follows that

$$\rho_{12} \leq \exp(-1/2) = 0.6065.$$

Similarly the probabilities for a given individual at times 2 and 3 differ by 1 on the logistic scale, from which follows the similar restriction $\rho_{23} \leq 0.6065$. Finally the probabilities for an individual at times 1 and 3 differ by 2, so that

$$\rho_{13} \leq \exp(-1) = 0.3679.$$

Hence $(\rho_{12}, \rho_{23}, \rho_{13}) = (0.3, 0.4, 0.2)$ are all within the admissible range for the intra-individual correlation parameters.

Finally, we assume a value of $\sigma^2 = 2.0$ for the random effects variance.

4.5 A Simulation Study

4.5.1 Results for Data With Covariates for Time Only

We initially simulated 300 data sets according to (4.4.1), modelling the response as a function of indicator variables for time only.

Consider first some graphical summaries of the results. The figures referred to in this and the following two subsections appear at the end of each respective subsection.

Plotted in figure 4.1 are the cumulative averages of the estimates of the fixed effects parameters, across the 300 data sets. Specifically, if $\hat{\beta}_{tv}$ is the estimate of $\hat{\beta}_t$ obtained from

the vth simulated data set, the graph shows

$$\frac{\sum_{v=1}^{r}\hat{\beta}_{tv}}{r} \quad \text{plotted against} \quad r, \quad t = 1, 2, 3, \quad r = 1, ..., 300.$$

Also plotted at regular intervals are points indicating one standard error above and below the current estimate of the cumulative average. These standard errors were computed as

$$\text{s.e.}\left(\frac{\sum_{v=1}^{r}\hat{\beta}_{tv}}{r}\right) = \sqrt{\frac{1}{r^2}\sum_{v=1}^{r}\widehat{\text{Var}}(\hat{\beta}_{tv})}.$$

The variability in the cumulative average of the estimates diminishes as more estimates enter the calculations. As a reference, horizontal lines are indicated at the true parameter values ($\beta_1 = -2$, $\beta_2 = -1$, $\beta_3 = 0$). We see that the averages of all three estimates settle somewhat higher than their true value, suggesting a slight bias.

Figure 4.2 shows similar plots for the intra-individual correlations. (Since no model-based estimate of the variance of $\rho_{tt'}$ is available, we used the sample standard error of the cumulative average of the estimates to compute standard errors as above). We witness considerable bias in the estimates of ρ_{12}, ρ_{23} and ρ_{13}; in all cases they underestimate the true value of the parameter. This however is not surprising. The true value of $\rho_{tt'}$ is conditional upon an unobserved set of random effects, whereas its estimate is computed conditional on the estimates of these random effects. But the random effects estimates, in adjusting for the overdispersion between schools, will in the process tend to adjust for some of the correlation between observations on the same individual as well. Similarly, if we were to ignore the clustering due to schools and estimate only fixed effects and intra-individual correlations, we would expect the correlation estimates to be larger than nominal, since these would then also be capturing some of the extraneous variability in the data due to the cross-sectional clustering. In the following two sections we shall see that this is indeed the case.

Figure 4.3 is a cumulative average plot of the random effects variance estimates. It seems that $\hat{\sigma^2}$ somewhat underestimates the true value, although the bias is not severe. Since for each simulated data set we first needed to generate a sample of random effects, we actually know for such data the true values of the "unobserved" random effects. From these an unbiased estimate of the true random effects variance (simply the sample variance) can be computed for each data set. A scatterplot of these sample variances versus the quasi empirical Bayes estimates is shown in figure 4.4. The 45 degree line through the origin is superimposed for reference. The points appear to be scattered roughly about this line, with the sample variances tending on average to be slightly larger than the corresponding model estimates, as anticipated.

Finally, we consider normal probability plots for the fixed effects parameter estimates. These are displayed in figure 4.5. The pronounced linear pattern in each of these plots validates the assumption of asymptotic normality for the estimates and hence justifies the use of standard normal confidence intervals.

Table 4.1 gives a numerical summary of the results. Shown are the averages of the parameter estimates across the 300 data sets, the averages of the "naive" or model-based standard errors, computed from $\hat{I}^{-1}_{QEB}$, as well as those of the robust standard errors, computed from $\hat{I}^{-1}_{QB}$. For comparison, the sample standard errors of the parameter estimates across all data sets are also included. Note that the robust standard errors are only slightly larger on average than the model-based ones; (the same will be observed in the next two sections and is pursued further there). This is not surprising, as it turns out that the parameter estimates from the quasi empirical Bayes model tend not to be sensitive to small changes in the prior variance on which they are conditioned (see sections 4.5.2 and 4.5.3). Both the model-based and the robust standard errors seem to underestimate the true variability in the parameter estimates, as a comparison with the sample standard errors indicates. The discrepancy is not large, however; observe also the similarity between the mean of the robust standard error estimate for σ^2 and the corresponding sample standard

Par.	True Value	Mean Est.	Mean s.e.(N)	Mean s.e.(R)	Sample s.e.
β_1	-2.0	-1.9320	.3718	.3764	.3913
β_2	-1.0	-0.9314	.3484	.3506	.3672
β_3	0.0	0.0538	.3414	.3422	.3616
ρ_{12}	0.3	0.2163	- -	- -	.0802
ρ_{23}	0.4	0.3154	- -	- -	.0681
ρ_{13}	0.2	0.1358	- -	- -	.0627
σ^2	2.0	1.8554	- -	.7797	.8392

Table 4.1: Numerical results of model fitting to simulated data, assuming covariates for time only.

error.

In tables 4.2 and 4.3 we consider confidence intervals for the fixed effects parameter estimates and investigate the empirical coverage rates for intervals of varying sizes. Table 4.2 reports coverage rates for the model-based confidence intervals. The entries in this table correspond to the number of datasets (out of 300) for which the true parameter value was included in the interval $\hat{\beta} \pm Z_q \cdot \text{s.e.}_N(\hat{\beta})$, where Z_q is the appropriate normal quantile, and the standard error is the model-based one for the particular estimate $\hat{\beta}$. The values in brackets express the coverage as a percentage. Table 4.3 reports coverage rates for the corresponding robust intervals $\hat{\beta} \pm Z_q \cdot \text{s.e.}_R(\hat{\beta})$. Neither the model-based nor the robust confidence intervals fully achieve the nominal coverage, though they are quite close. The robust intervals improve somewhat on the model-based ones. In interpreting such empirical coverage rates one must also keep in mind the magnitude of the error to be expected, due to performing only a finite number of simulations. For tables 4.2 and 4.3, computing the standard deviation of the estimate $\hat{p}_\alpha$ of a nominal coverage rate p_α as $\sqrt{p_\alpha(1-p_\alpha)/300}$, we find that most of the observed coverage rates lie within 3 standard deviations of the

Nominal Coverage	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
99 %	297 (0.990)	294 (0.980)	290 (0.967)
98 %	295 (0.983)	288 (0.960)	287 (0.957)
95 %	278 (0.927)	277 (0.923)	275 (0.917)
90 %	259 (0.863)	253 (0.843)	253 (0.843)
80 %	230 (0.767)	228 (0.760)	228 (0.760)

Table 4.2: Coverage rates for model-based confidence intervals, assuming covariates for time only.

Nominal Coverage	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
99 %	297 (0.990)	294 (0.980)	290 (0.967)
98 %	296 (0.987)	290 (0.967)	289 (0.963)
95 %	278 (0.927)	278 (0.927)	275 (0.917)
90 %	259 (0.863)	253 (0.843)	254 (0.847)
80 %	231 (0.770)	228 (0.760)	228 (0.760)

Table 4.3: Coverage rates for robust confidence intervals, assuming covariates for time only.

corresponding nominal value. In the analogous tables in sections 4.5.2 and 4.5.3, they all lie within this range of the nominal value. (Using $\hat{p}_\alpha \pm 3s.d.(\hat{p}_\alpha)$ as a plausible range for an estimate of p_α relies on the normal approximation to the binomial, which is reasonable as long as the smaller of np_α and nq_α is greater than 5, where n is the sample size and $q_\alpha = 1 - p_\alpha$. This condition is satisfied for most coverage rates given in tables 4.2 and 4.3 and in analogous tables in the subsequent sections).

A similar but more extensive examination of the performance of the quasi empirical Bayes model is given in the next two sections, where we fit the model to data with either an individual or a cluster-level covariate, and compare the results to those of other analyses.

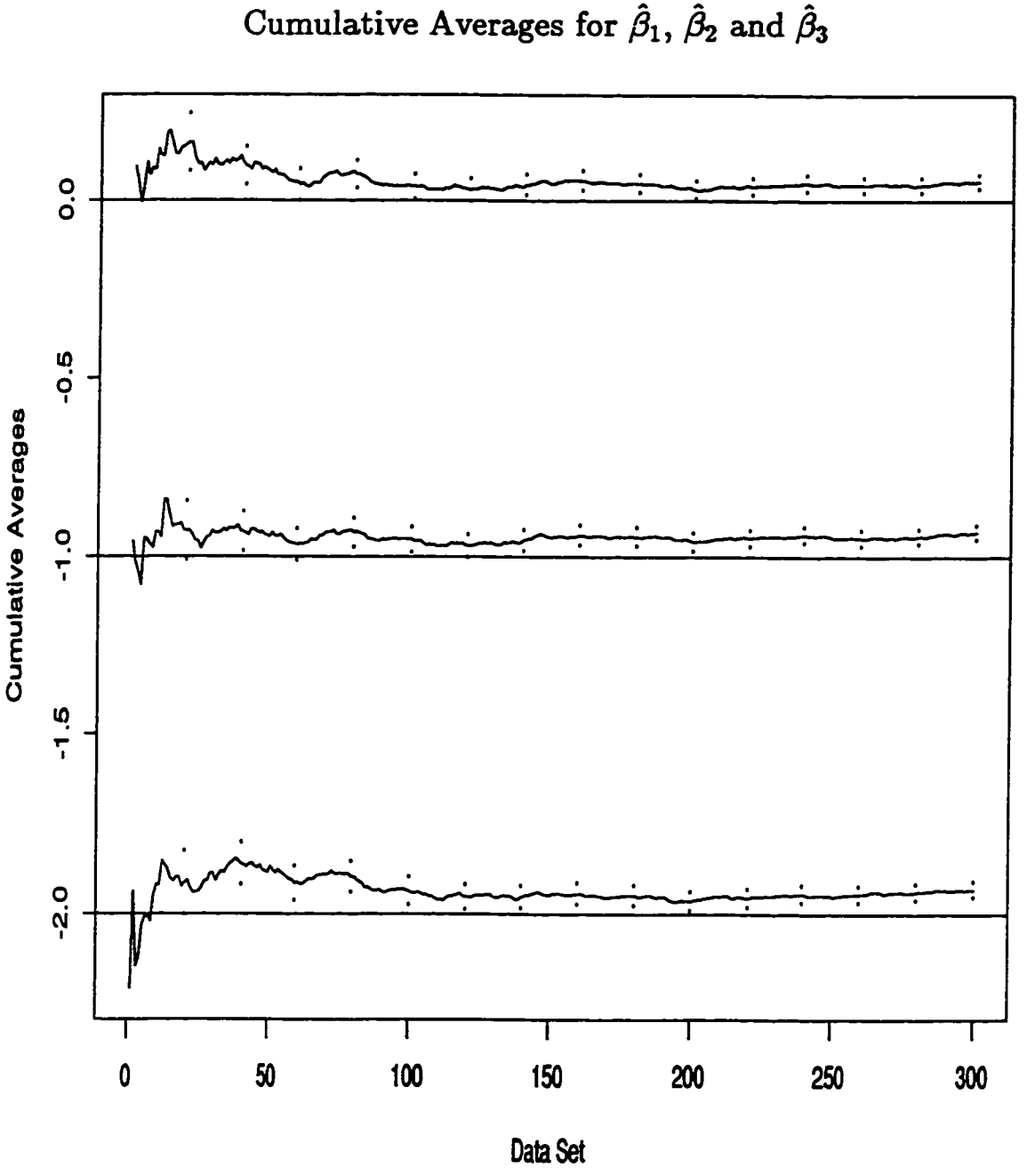


Figure 4.1: Cumulative averages of fixed effects parameter estimates ($\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$), assuming covariates for time only; points indicate ± 1 std. error.

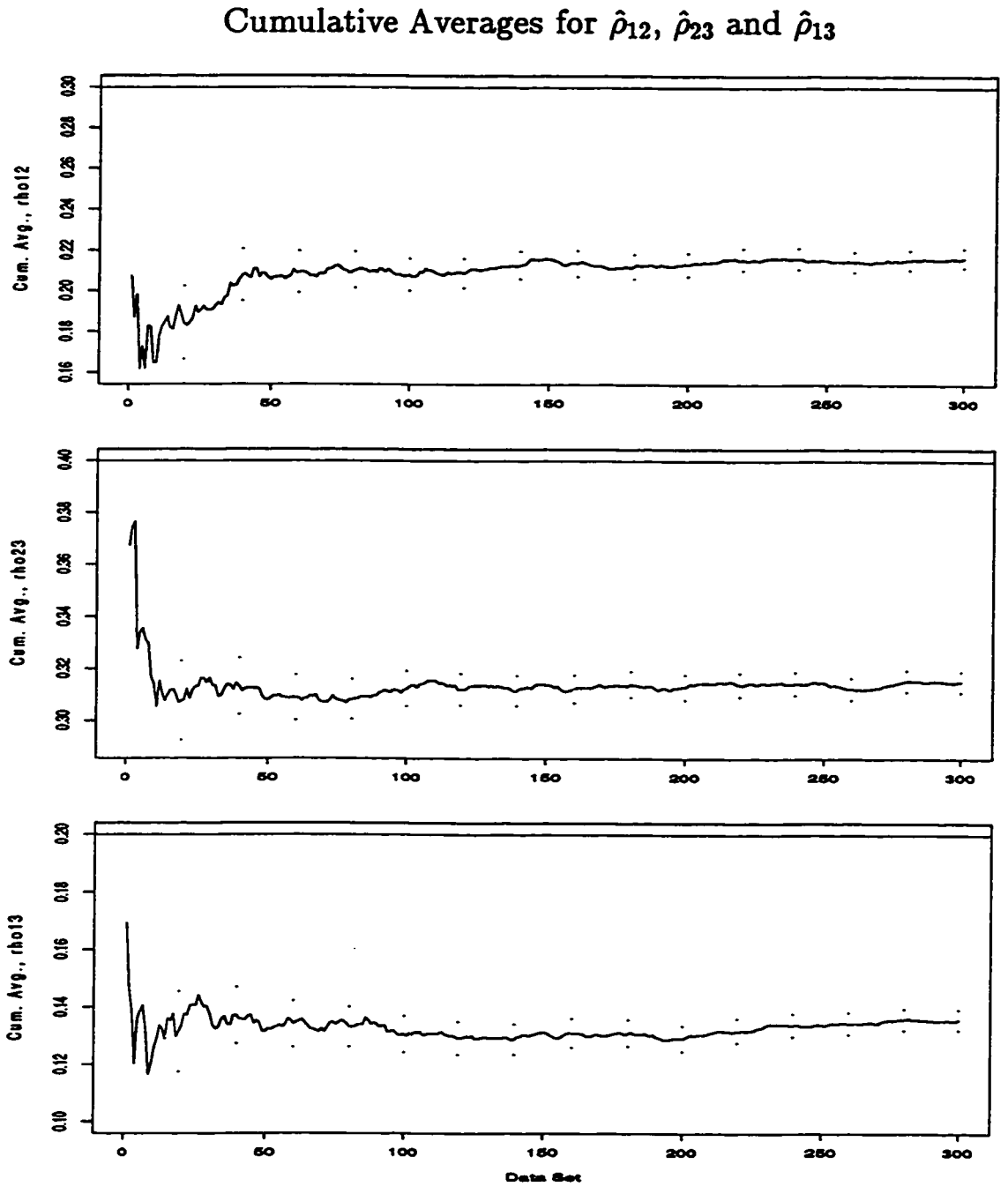


Figure 4.2: Cumulative averages of intra-individual correlation estimates ($\hat{\rho}_{12}$, $\hat{\rho}_{23}$, $\hat{\rho}_{13}$), assuming covariates for time only; points indicate ± 1 std. error.

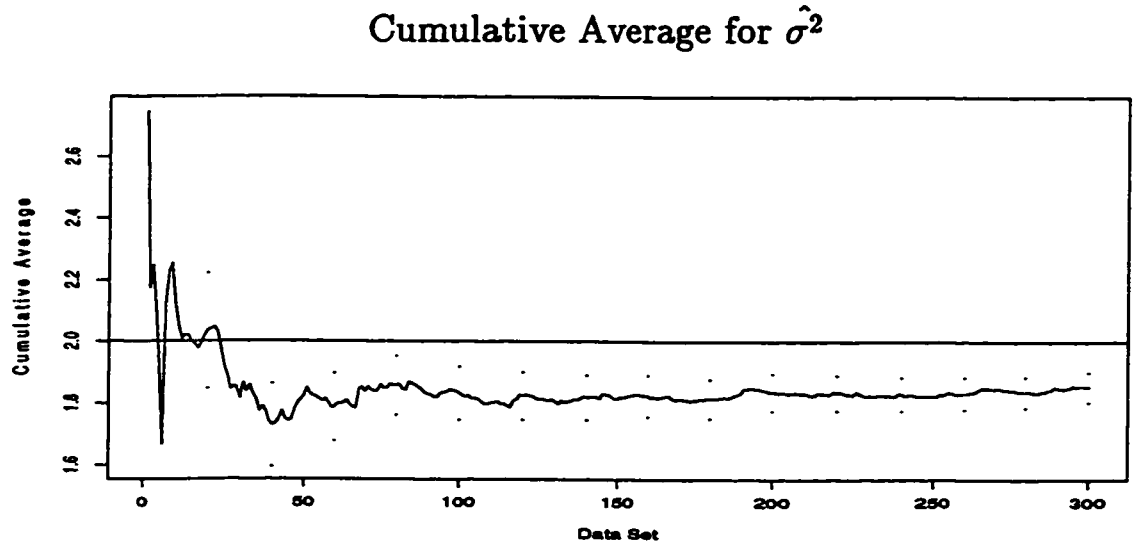


Figure 4.3: Cumulative average of prior variance estimates ($\hat{\sigma^2}$), assuming covariates for time only; points indicate ± 1 std. error.

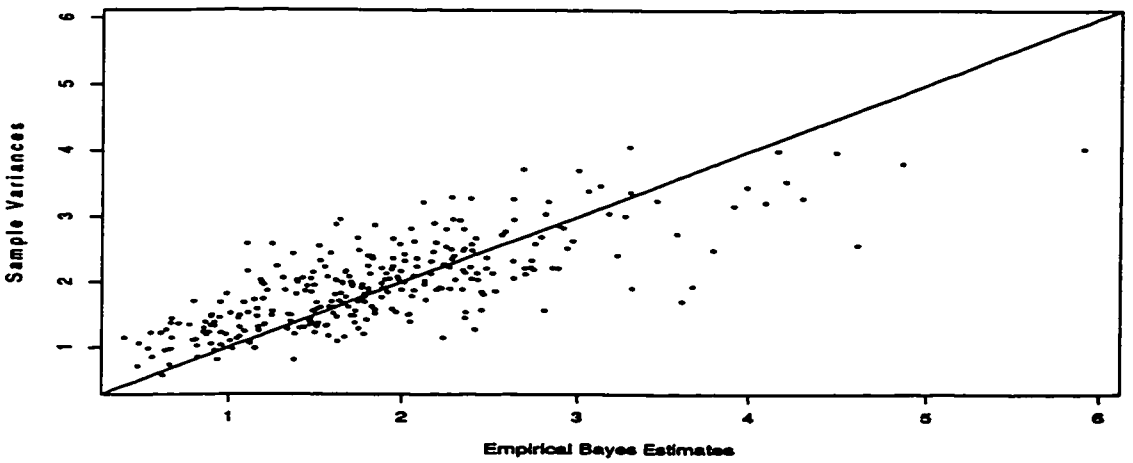


Figure 4.4: QEB Estimates of the random effects variance versus sample variances of the (simulated) random effects, assuming covariates for time only.

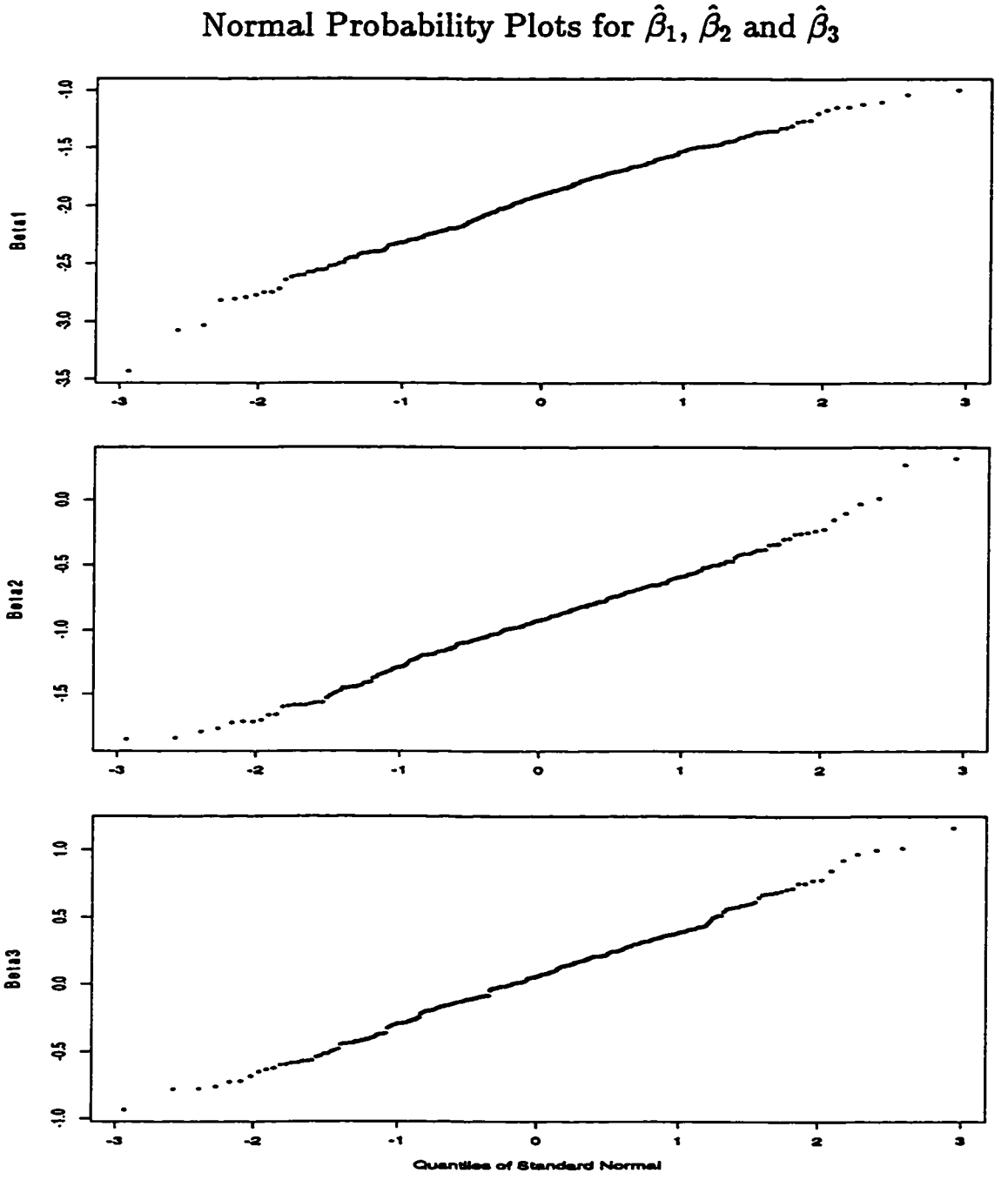


Figure 4.5: Normal probability plots for fixed effects parameter estimates ($\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$), assuming covariates for time only.

4.5.2 Results for Data With an Individual-Level Covariate

We next simulated 300 data sets according to (4.4.1), modelling the response as a function of indicators for time, as well as an individual-level covariate (see the description in section A 2 of the Appendix to this chapter).

As in the previous section, figure 4.6 shows plots of the cumulative averages of the estimates of the fixed effects parameters across the 300 data sets. Here we do not observe the cumulative averages to be consistently larger or smaller than the true parameter values, and for none of the parameters does the overall mean of the estimates appear to be significantly different from the true value. The bias in the fixed effects parameter estimates therefore seems to be negligible.

Considering similar plots for the intra-individual correlations (see figure 4.7), we make the same observations as before. The estimates of ρ_{12}, ρ_{23} and ρ_{13} underestimate the true values of the correlation parameters.

Figure 4.8 shows the cumulative average plot of the random effects variance estimates. As before it appears that $\hat{\sigma^2}$ slightly underestimates the true value, but this bias seems insignificant. A scatterplot of sample variances versus the quasi empirical Bayes estimates is given in figure 4.9.

The normal plots for the fixed effects parameters are displayed in figure 4.10. Again we detect no significant departure from linearity in any of the graphs.

Table 4.4 gives a numerical summary of the results for this set of simulations, similar to that in table 4.1. As seen from the graphs referred to above, the fixed effects parameter estimates seem to be virtually unbiased. Note that here the model-based, robust and sample standard errors are all roughly the same size, so that confidence interval coverage properties can be expected to be quite good. In table 4.5 we give the empirical coverage rates for $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$, for model-based confidence intervals; in table 4.6 coverages for the corresponding robust intervals are given. Indeed the results in both cases are auspi-

Par.	True Value	Mean Est.	Mean s.e.(N)	Mean s.e.(R)	Sample s.e.
β_1	-2.0	-1.9873	.3949	.3989	.3957
β_2	-1.0	-1.0065	.3765	.3780	.3804
β_3	0.0	-0.0227	.3702	.3705	.3712
β_4	1.0	0.9997	.2456	.2474	.2662
ρ_{12}	0.3	0.2233	- -	- -	.0817
ρ_{23}	0.4	0.3134	- -	- -	.0738
ρ_{13}	0.2	0.1379	- -	- -	.0648
σ^2	2.0	1.9381	- -	.7931	.7635

Table 4.4: Numerical results of model fitting to simulated data, assuming inclusion of an individual-level covariate.

cious; the nominal coverage rates are almost attained, especially by the robust confidence intervals.

We now examine the fit of the quasi empirical Bayes model more closely for several particular cases. Out of the 300 simulated data sets consider specifically the five data sets giving estimates of the prior variance σ^2 corresponding to the largest and smallest of the 300, the endpoints of the interquartile range and the median. We first wish to investigate the sensitivity of the results to changes in the value of σ^2. For this purpose we computed for each of these five data sets the standard error of $\hat{\sigma^2} = \hat{\sigma^2}_{QBB}$ (on the basis of I_{QB}^{-1}), and refit the model twice, taking $(\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2}))$ and $(\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2}))$ in turn as the 'true' fixed value of the random effects variance. (Note that this refitting should not be referred to as quasi *empirical* Bayes, since no updating of the variance parameter is involved). The results are tabulated in table 4.7. For $\hat{\sigma^2} = 0.508$, 1.383, 1.782, 2.429 and 4.385 the estimates of β_1, β_2, β_3 and β_4 are shown, along with their (model-based) standard errors, assuming that $\sigma^2 = \hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$, $\sigma^2 = \hat{\sigma^2}$ and $\sigma^2 = \hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$. Given the relative insensitivity of the

Nominal Coverage	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
99 %	298 (0.993)	294 (0.980)	298 (0.993)	296 (0.987)
98 %	293 (0.977)	292 (0.973)	294 (0.980)	291 (0.970)
95 %	282 (0.940)	276 (0.920)	283 (0.943)	281 (0.937)
90 %	269 (0.897)	267 (0.890)	267 (0.890)	257 (0.857)
80 %	242 (0.807)	245 (0.817)	236 (0.787)	229 (0.763)

Table 4.5: Coverage rates for model-based confidence intervals, assuming inclusion of an individual-level covariate.

Nominal Coverage	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
99 %	298 (0.993)	294 (0.980)	298 (0.993)	296 (0.987)
98 %	293 (0.977)	292 (0.973)	294 (0.980)	292 (0.973)
95 %	283 (0.943)	278 (0.927)	284 (0.947)	282 (0.940)
90 %	269 (0.897)	267 (0.890)	267 (0.890)	260 (0.867)
80 %	246 (0.820)	245 (0.817)	235 (0.783)	229 (0.763)

Table 4.6: Coverage rates for robust confidence intervals, assuming inclusion of an individual-level covariate.

estimates of β to even substantial changes in σ^2, as seen in table 4.7, it makes sense that the robust standard errors for these parameters, taking into account the variability due to the fact that σ^2 has to be estimated, are only marginally larger than the ones from the model.

Consider now a comparison of the quasi empirical Bayes model with three other models, each a special case of this more general model: to each of the five data sets described above we fit a standard logistic model, a standard logistic-normal empirical Bayes model addressing the cross-sectional clustering but not the longitudinal component, a GEE model accomodating the longitudinal nature of the data but not the cross-sectional, and finally

	σ^2	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	s.e. $(\hat{\beta}_1)$	s.e. $(\hat{\beta}_2)$	s.e. $(\hat{\beta}_3)$	s.e. $(\hat{\beta}_4)$
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	0.102	-1.865	-1.057	-0.058	0.579	.243	.214	.200	.239
$\hat{\sigma^2}$	0.508	-1.963	-1.123	-0.064	0.606	.289	.263	.251	.242
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	0.914	-2.007	-1.155	-0.072	0.620	.322	.299	.289	.245
s.e.$(\hat{\sigma^2})$:	0.406								
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	0.756	-1.637	-0.827	0.175	0.812	.298	.282	.276	.241
$\hat{\sigma^2}$	1.383	-1.702	-0.858	0.181	0.854	.351	.336	.331	.247
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	2.006	-1.736	-0.875	0.184	0.876	.395	.382	.376	.251
s.e.$(\hat{\sigma^2})$:	0.623								
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	1.034	-2.268	-1.486	-0.537	1.228	.339	.319	.306	.244
$\hat{\sigma^2}$	1.782	-2.363	-1.556	-0.571	1.278	.396	.377	.365	.250
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	2.531	-2.418	-1.598	-0.592	1.305	.444	.427	.415	.253
s.e.$(\hat{\sigma^2})$:	0.748								
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	1.440	-1.512	-0.532	0.385	0.678	.351	.339	.339	.244
$\hat{\sigma^2}$	2.429	-1.582	-0.571	0.380	0.711	.420	.409	.410	.250
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	3.418	-1.622	-0.595	0.375	0.730	.478	.468	.468	.253
s.e.$(\hat{\sigma^2})$:	0.989								
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	2.869	-1.906	-1.038	0.164	1.236	.461	.449	.442	.263
$\hat{\sigma^2}$	4.385	-1.975	-1.072	0.171	1.292	.542	.531	.523	.271
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	5.900	-2.015	-1.091	0.176	1.325	.611	.600	.593	.276
s.e.$(\hat{\sigma^2})$:	1.515								

Table 4.7: Sensitivity of fixed effects estimates and standard errors to the prior variance σ^2, assuming inclusion of an individual-level covariate; (each horizontal panel corresponds to one data set).

the composite quasi empirical Bayes model. The results are shown in table 4.8.

The (standard) empirical Bayes model improves on the logistic in that it more accurately reflects the variability in the estimates of the intercept parameters β_1, β_2 and β_3; as the variance of the random effects (i.e. the overdispersion between schools) increases, so does the variance of these estimates, as one might expect. However this model does not show a substantial inflation in the variance of the estimate of β_4, the parameter for the individual-level covariate, as compared to the logistic fit. This is due to the fact that standard empirical Bayes neglects to account for the intra-individual correlations. Note also that since the estimation of β_4 is confined to comparisons between individuals *within* schools, its variability is largely unaffected by the size of σ^2.

In contrast, the GEE model improves on the logistic in that by estimating the intra-individual correlations, it more accurately reflects the variability in $\hat{\beta}_4$, as one would anticipate for an individual-level covariate. The increase in the standard error for this parameter over that estimated from the logistic model is roughly constant over the given range of σ^2. The drawback with GEE is the fact that because it ignores the cross-sectional overdispersion, the standard errors of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ are in each case understated (roughly in keeping with the estimates of the logistic model).

The quasi empirical Bayes model is the only one which properly adjusts for both forms of clustering in the data. The nominal standard errors of the time parameter estimates are duly inflated, as in the empirical Bayes fit, and that of the estimate for the individual-level covariate is also adjusted, as in the GEE fit.

Note also that those approaches ignoring the presence of random effects (the logistic and GEE models) produce coefficient estimates which are attenuated from the ones given by the empirical or quasi empirical Bayes fits. These purely population-averaged approaches effectively average over the random effects distribution in computing parameter estimates. The resulting attenuation factor on the coefficients, comparing with the estimates from a cluster-specific model, depends on the random effects variance and approximately equals

	MODELS:							
Par.	Logistic		E B		G E E		Q E B	
β_1	-1.821	(.214)	-1.974	(.286)	-1.835	(.234)	-1.963	(.292)
β_2	-1.024	(.182)	-1.123	(.259)	-1.039	(.203)	-1.123	(.264)
β_3	-0.0543	(.169)	-0.0625	(.247)	-0.0671	(.189)	-0.0636	(.249)
β_4	0.559	(.183)	0.605	(.191)	0.577	(.246)	0.606	(.242)
ρ_{12}					0.432		0.381	
ρ_{23}					0.486		0.397	
ρ_{13}					0.283		0.209	
σ^2			0.590	(.315)			0.508	(.406)
β_1	-1.385	(.191)	-1.724	(.361)	-1.385	(.213)	-1.702	(.353)
β_2	-0.712	(.173)	-0.871	(.346)	-0.712	(.195)	-0.858	(.337)
β_3	0.122	(.168)	0.180	(.340)	0.122	(.189)	0.181	(.330)
β_4	0.675	(.175)	0.851	(.200)	0.676	(.247)	0.854	(.249)
ρ_{12}					0.530		0.333	
ρ_{23}					0.537		0.373	
ρ_{13}					0.405		0.255	
σ^2			1.596	(.624)			1.383	(.623)
β_1	-1.849	(.211)	-2.438	(.411)	-1.805	(.228)	-2.363	(.402)
β_2	-1.216	(.187)	-1.624	(.390)	-1.174	(.207)	-1.556	(.381)
β_3	-0.467	(.172)	-0.625	(.376)	-0.428	(.193)	-0.571	(.366)
β_4	1.020	(.184)	1.338	(.218)	0.991	(.248)	1.278	(.252)
ρ_{12}					0.389		0.193	
ρ_{23}					0.492		0.221	
ρ_{13}					0.429		0.228	
σ^2			2.007	(.774)			1.782	(.748)
β_1	-1.149	(.183)	-1.621	(.426)	-1.135	(.204)	-1.582	(.422)
β_2	-0.385	(.168)	-0.615	(.415)	-0.368	(.190)	-0.571	(.410)
β_3	0.300	(.168)	0.349	(.415)	0.316	(.190)	0.380	(.409)
β_4	0.526	(.172)	0.718	(.203)	0.514	(.246)	0.711	(.251)
ρ_{12}					0.542		0.297	
ρ_{23}					0.576		0.372	
ρ_{13}					0.429		0.197	
σ^2			2.598	(.977)			2.429	(.989)

(continued)

	MODELS:							
Par.	Logistic		E B		G E E		Q E B	
β_1	-1.171	(.183)	-1.963	(.545)	-1.166	(.207)	-1.975	(.544)
β_2	-0.660	(.171)	-1.055	(.532)	-0.655	(.195)	-1.072	(.531)
β_3	0.0505	(.167)	0.180	(.525)	0.0562	(.190)	0.172	(.523)
β_4	0.738	(.172)	1.289	(.239)	0.730	(.251)	1.292	(.274)
ρ_{12}					0.601		0.164	
ρ_{23}					0.601		0.233	
ρ_{13}					0.462		0.153	
σ^2			4.488	(1.502)			4.385	(1.515)

Table 4.8: Comparison of the QEB model with the logistic, empirical Bayes (EB), and GEE fits, assuming inclusion of an individual-level covariate; (standard errors are indicated in brackets - each horizontal panel corresponds to one data set).

$(1+c^2\sigma^2)^{-1/2}$, where $c = 16\sqrt{3}/15\pi$; see section 3.5.6 and equation (3.5.23). Observe that the quasi empirical Bayes model is neither purely a population-averaged, nor an entirely cluster-specific approach. It has a population-averaged interpretation insofar as it models in a marginal sense the repeated observations on individuals within schools, and a cluster-specific interpretation insofar as it conditions on random effects for schools.

In the previous section we gave a brief justification of why the intra-individual correlations, as estimated from the quasi empirical Bayes model, will tend to somewhat underestimate the true values of these parameters. If we consider the GEE correlation estimates given in table 4.8, however, we see that these seem to overstate the correlations. As indicated, this is due to some of the extraneous variability from the cross-sectional overdispersion being captured in $\hat{\rho}_{tt'}$. As a consequence, neither the QEB nor the GEE estimates of correlation are unbiased, but note that the former at least are conditioned on estimates of the school random effects, whereas the latter are not. The QEB estimates will therefore give a better indication of the true values of the conditional correlations

$\text{Corr}(Y_{kjt}, Y_{kjt'}|b_k) = \rho_{tt'}$ than the GEE estimates.

In the next section we consider one more set of simulations, to investigate the performance of the quasi empirical Bayes model on data with a cluster-level covariate.

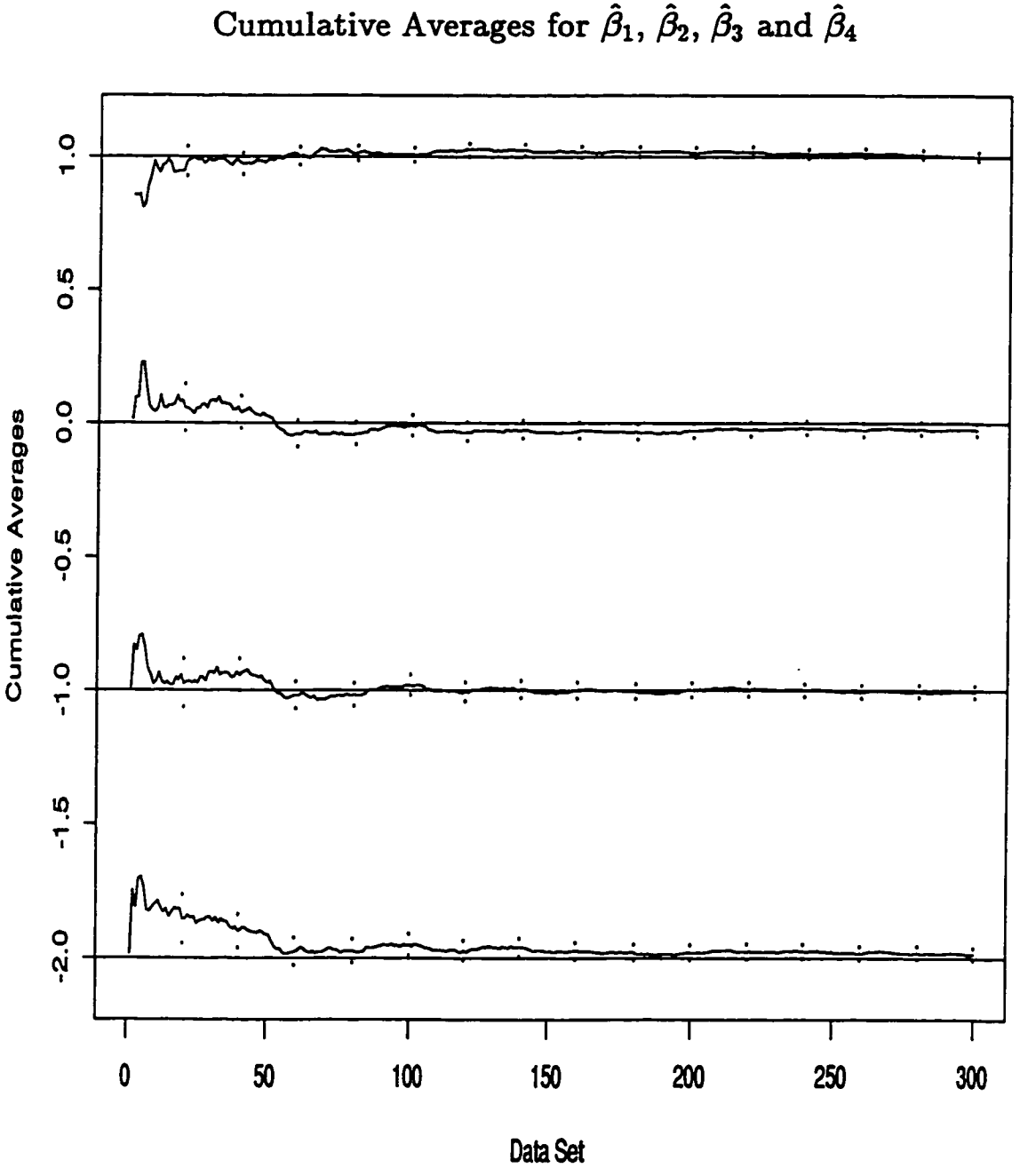


Figure 4.6: Cumulative averages of estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, assuming inclusion of an individual-level covariate; points indicate ± 1 std. error.

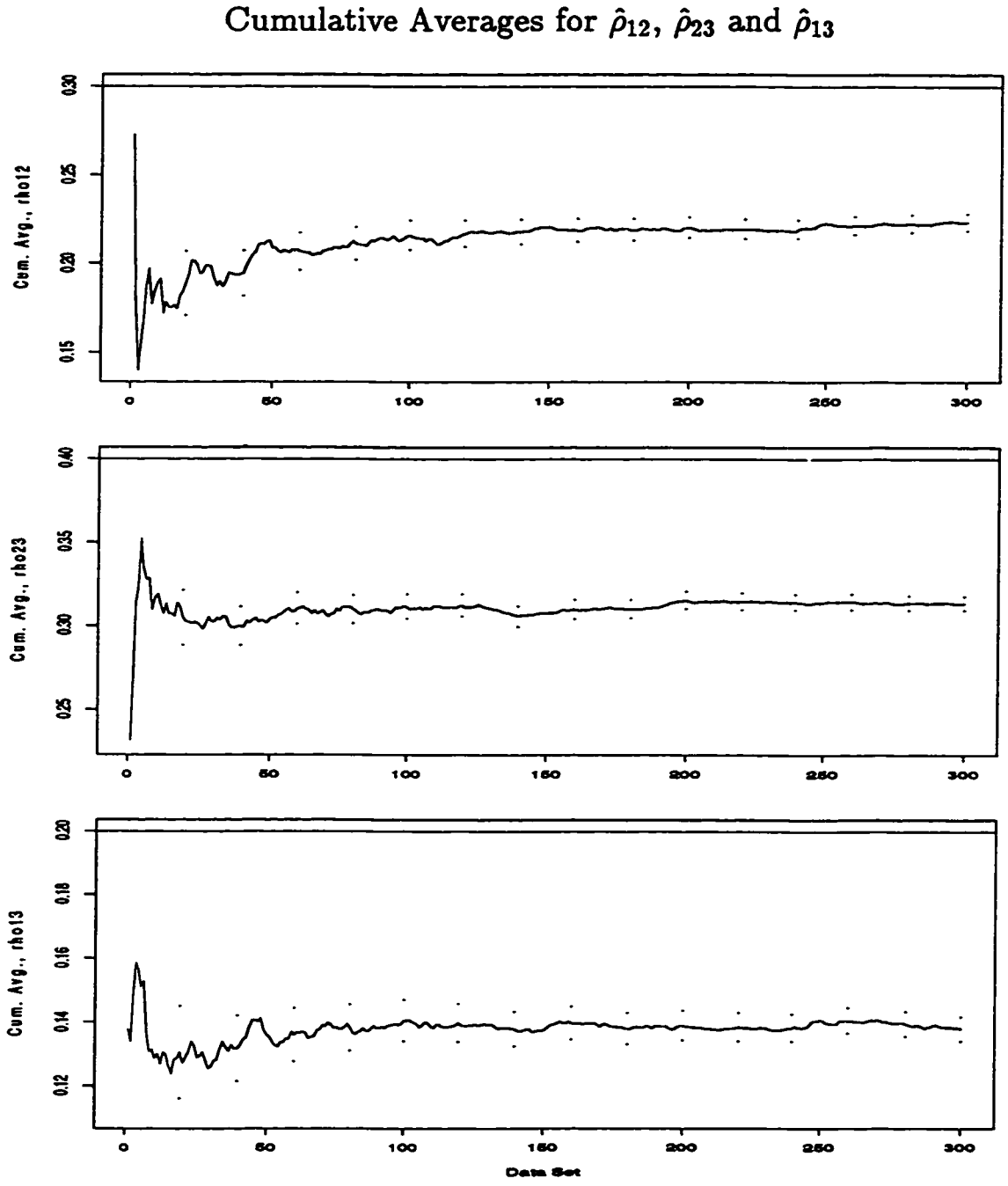


Figure 4.7: Cumulative averages of correlation estimates $\hat{\rho}_{12}$, $\hat{\rho}_{23}$, $\hat{\rho}_{13}$, assuming inclusion of an individual-level covariate; points indicate ± 1 std. error.

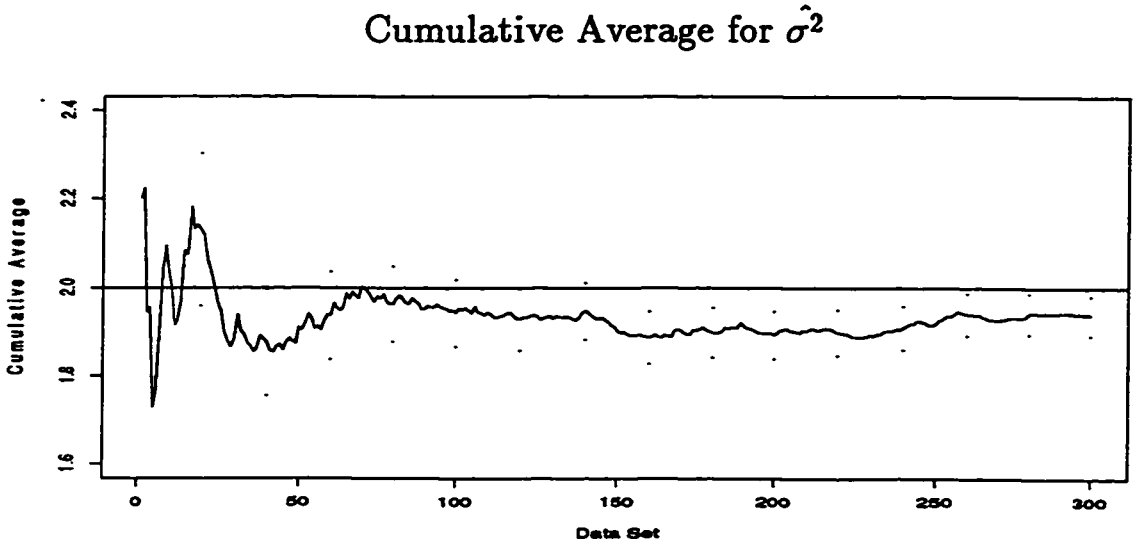


Figure 4.8: Cumulative average of variance estimates $\hat{\sigma^2}$, assuming inclusion of an individual-level covariate; points indicate ± 1 std. error.

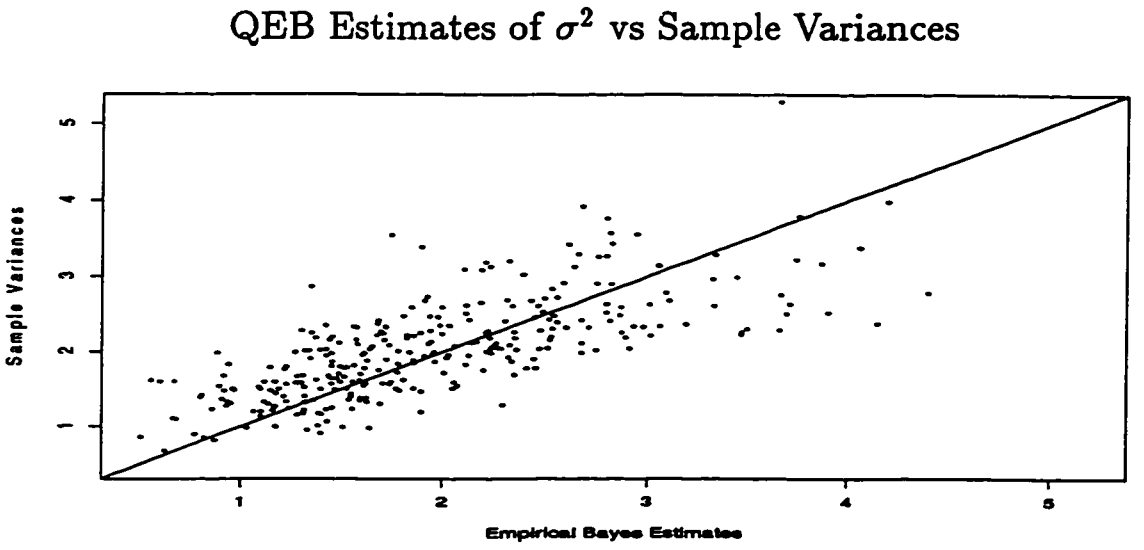


Figure 4.9: QEB Estimates of σ^2 versus sample variances of the (simulated) random effects, assuming inclusion of an individual-level covariate.

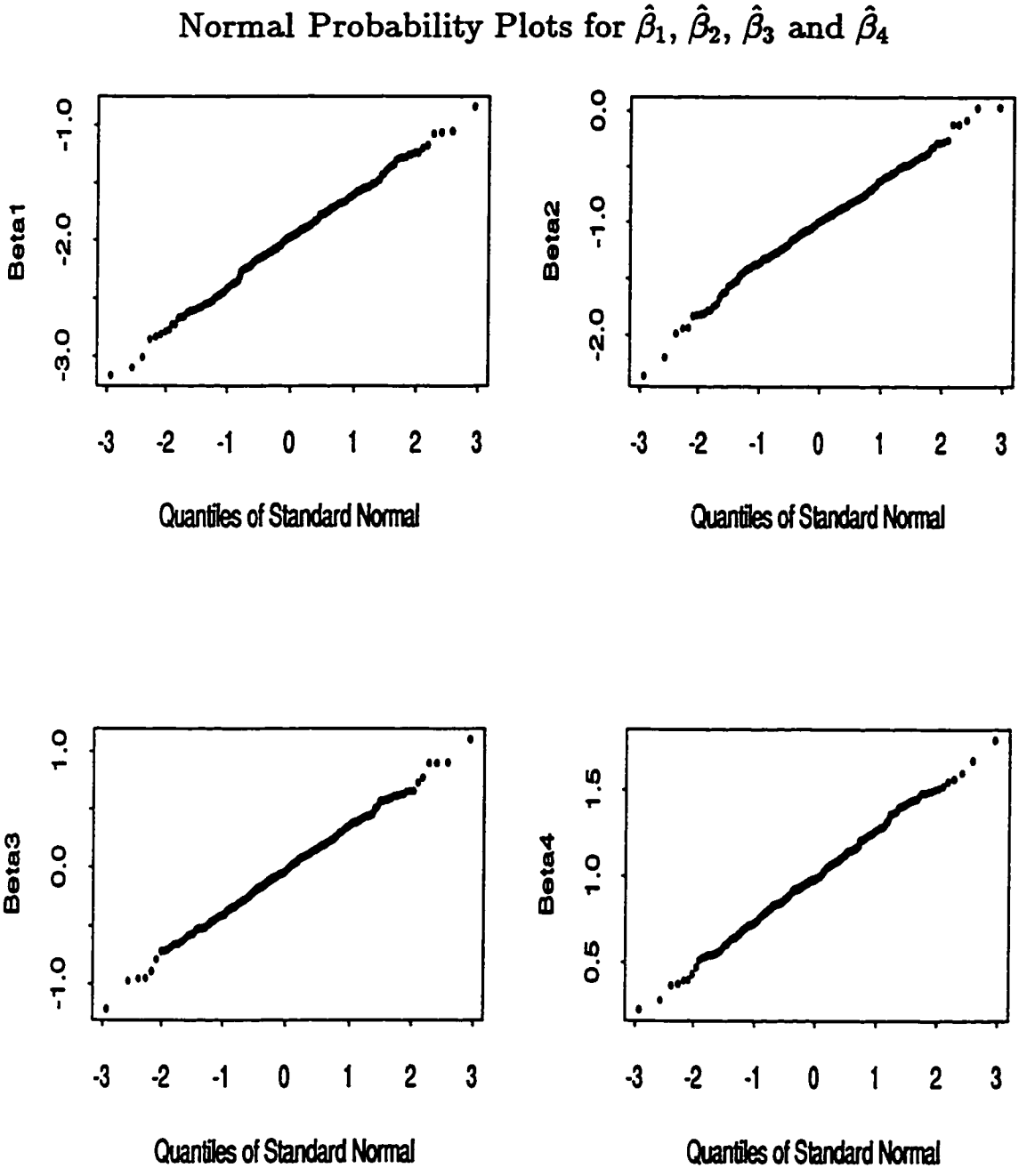


Figure 4.10: Normal probability plots for estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$, assuming inclusion of an individual-level covariate.

4.5.3 Results for Data With a Cluster-Level Covariate

We simulated 300 data sets according to (4.4.1), modelling the response as a function of indicators for time, as well as a cluster- or school-level covariate (see the description in section A 3 of the Appendix to this chapter).

Figure 4.11 shows plots of the cumulative averages of the estimates of the fixed effects parameters across the 300 data sets. The intercept parameter estimates are again slightly larger on average than the true parameter values, whereas the coefficient for the cluster-level covariate is somewhat underestimated. The intra-individual correlation estimates behave as in the previous two sections (see figure 4.12). Considering the cumulative average plot of the random effects variance estimates (figure 4.13), these estimates seem to settle exactly at the true value of the prior variance. The scatterplot of sample variances versus the quasi empirical Bayes estimates suggests, as in the previous two sections, that $\hat{\sigma^2}$ underestimates σ^2 for smaller values of the prior variance, but overestimates σ^2 for larger values. Here a balance seems to have been achieved, in view of the average of the 300 estimates.

Par.	True Value	Mean Est.	Mean s.e.(N)	Mean s.e.(R)	Sample s.e.
β_1	-2.0	-1.9158	.5059	.5089	.4961
β_2	-1.0	-0.9261	.4927	.4939	.4871
β_3	0.0	0.0682	.4884	.4887	.4638
β_4	1.0	0.9216	.6719	.6729	.6948
ρ_{12}	0.3	0.2260	- -	- -	.0771
ρ_{23}	0.4	0.3187	- -	- -	.0760
ρ_{13}	0.2	0.1416	- -	- -	.0655
σ^2	2.0	1.9947	- -	.8292	.7961

Table 4.9: Numerical results of model fitting to simulated data, assuming inclusion of a cluster-level covariate.

Table 4.9 gives a numerical summary of the results for this set of simulations, similar to those in the previous two sections. Curiously the sample standard errors for the intercept parameters β_1, β_2 and β_3 turned out to be slightly smaller than the means of the model-based and robust standard errors, whereas the sample standard error for β_4 is larger. Comparing robust and sample standard errors for σ^2, we find that they agree fairly well, as in the previous two sections.

In table 4.10 we give the empirical coverage rates for $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$, for model-based confidence intervals; in table 4.11 coverages for the corresponding robust intervals are given. As expected, the coverage rates for $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ are excellent; however they are a bit understated for $\hat{\beta}_4$. The robust intervals closely resemble the model-based ones.

As in the last section we next examine the fit of the quasi empirical Bayes model more closely for the five specific data sets yielding the smallest and largest estimates of σ^2, and the estimates delimiting the interquartile range and the median. In table 4.12 the sensitivity of the results to changes in the value of σ^2 is investigated, in the same manner as in table 4.7. The sample of random effects variance estimates in this case was $\{0.477, 1.379, 1.930, 2.446, 4.479\}$. Again we observe that whereas the standard errors of the parameter estimates do increase or decrease fairly substantially according to changes in the value of σ^2, the point estimates themselves, in comparison, remain relatively insensitive to such changes.

As before, consider now a comparison of the quasi empirical Bayes model including a cluster-level covariate with the standard logistic, empirical Bayes and GEE models. Referring to table 4.13, we note the similarity in point estimates and standard errors between the empirical Bayes and the quasi empirical Bayes analyses. Here the standard error of $\hat{\beta}_4$ is a function of the size of the random effects variance, and increases with larger values of σ^2. One would expect to see such behaviour, bearing in mind that β_4 is now the coefficient of a cluster-level covariate (i.e. the effect of this covariate is estimated in terms of a comparison between schools, and is therefore sensitive to cross-sectional overdispersion). The estimate

Nominal Coverage	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
99 %	297 (0.990)	299 (0.997)	297 (0.990)	294 (0.980)
98 %	293 (0.977)	295 (0.983)	295 (0.983)	291 (0.970)
95 %	281 (0.937)	283 (0.943)	289 (0.963)	277 (0.923)
90 %	270 (0.900)	267 (0.890)	274 (0.913)	260 (0.867)
80 %	238 (0.793)	229 (0.763)	241 (0.803)	224 (0.747)

Table 4.10: Coverage rates for model-based confidence intervals, assuming inclusion of a cluster-level covariate.

Nominal Coverage	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
99 %	297 (0.990)	299 (0.997)	297 (0.990)	294 (0.980)
98 %	293 (0.977)	295 (0.983)	295 (0.983)	292 (0.973)
95 %	283 (0.943)	284 (0.947)	289 (0.963)	277 (0.923)
90 %	270 (0.900)	266 (0.887)	274 (0.913)	260 (0.867)
80 %	238 (0.793)	229 (0.763)	241 (0.803)	224 (0.747)

Table 4.11: Coverage rates for robust confidence intervals, assuming inclusion of a cluster-level covariate.

of σ^2 and its standard error are also similar in the empirical Bayes and quasi empirical Bayes models. The similarity of these two analyses implies that if interest is only focussed on the parameters in the mean specification of model (4.4.1) and not the intra-individual correlations, and provided that only cluster-level covariates are to be investigated, then a standard empirical Bayes model can be applied to estimate the parameters (β, σ^2) from (4.4.1); the point estimates and their standard errors will closely reflect those from the quasi empirical Bayes model. This fact may be of advantage, since the empirical Bayes model is conceptually simpler than the quasi empirical Bayes, and also considerably easier to fit in a computational sense.

	σ^2	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	s.e. $(\hat{\beta}_1)$	s.e. $(\hat{\beta}_2)$	s.e. $(\hat{\beta}_3)$	s.e. $(\hat{\beta}_4)$
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	0.141	-1.486	-0.667	0.142	0.350	.241	.221	.216	.276
$\hat{\sigma^2}$	0.477	-1.546	-0.697	0.151	0.360	.308	.291	.288	.383
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	0.814	-1.575	-0.712	0.152	0.366	.358	.343	.340	.460
s.e.$(\hat{\sigma^2})$:	0.337								
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	0.728	-2.203	-1.026	0.024	1.742	.364	.344	.338	.456
$\hat{\sigma^2}$	1.379	-2.317	-1.107	-0.010	1.835	.451	.434	.428	.586
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	2.029	-2.383	-1.159	-0.036	1.891	.522	.507	.502	.691
s.e.$(\hat{\sigma^2})$:	0.651								
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	1.125	-2.225	-1.077	-0.305	1.540	.419	.402	.395	.538
$\hat{\sigma^2}$	1.930	-2.342	-1.151	-0.345	1.643	.511	.497	.490	.676
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	2.734	-2.415	-1.200	-0.375	1.712	.587	.575	.569	.789
s.e.$(\hat{\sigma^2})$:	0.804								
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	1.463	-2.311	-1.310	-0.283	1.583	.463	.447	.441	.599
$\hat{\sigma^2}$	2.446	-2.425	-1.395	-0.330	1.666	.566	.553	.547	.750
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	3.430	-2.496	-1.451	-0.366	1.721	.652	.640	.635	.875
s.e.$(\hat{\sigma^2})$:	0.984								
$\hat{\sigma^2} - \text{s.e.}(\hat{\sigma^2})$	2.906	-2.927	-1.540	-0.575	2.313	.625	.598	.590	.816
$\hat{\sigma^2}$	4.479	-3.038	-1.589	-0.597	2.429	.747	.722	.715	.996
$\hat{\sigma^2} + \text{s.e.}(\hat{\sigma^2})$	6.052	-3.105	-1.621	-0.614	2.504	.850	.827	.820	1.147
s.e.$(\hat{\sigma^2})$:	1.573								

Table 4.12: Sensitivity of fixed effects estimates and standard errors to the prior variance σ^2, assuming inclusion of a cluster-level covariate; (each horizontal panel corresponds to one data set).

	MODELS:							
Par.	Logistic		E B		G E E		Q E B	
β_1	-1.443	(.196)	-1.561	(.317)	-1.450	(.212)	-1.546	(.308)
β_2	-0.641	(.172)	-0.692	(.300)	-0.650	(.188)	-0.697	(.289)
β_3	0.135	(.167)	0.156	(.297)	0.127	(.183)	0.151	(.285)
β_4	0.338	(.176)	0.361	(.386)	0.346	(.228)	0.360	(.379)
ρ_{12}					0.285		0.184	
ρ_{23}					0.472		0.366	
ρ_{13}					0.233		0.134	
σ^2			0.574	(.295)			0.477	(.337)
β_1	-1.883	(.212)	-2.357	(.469)	-1.892	(.233)	-2.317	(.457)
β_2	-0.824	(.181)	-1.116	(.452)	-0.834	(.200)	-1.107	(.438)
β_3	0.0525	(.174)	-0.0283	(.446)	0.0515	(.190)	-0.00992	(.429)
β_4	1.511	(.191)	1.895	(.603)	1.508	(.255)	1.835	(.590)
ρ_{12}					0.452		0.372	
ρ_{23}					0.480		0.266	
ρ_{13}					0.255		0.103	
σ^2			1.546	(.653)			1.379	(.651)
β_1	-1.712	(.203)	-2.217	(.502)	-1.729	(.226)	-2.342	(.514)
β_2	-0.802	(.176)	-1.031	(.486)	-0.821	(.199)	-1.151	(.498)
β_3	-0.211	(.170)	-0.251	(.481)	-0.225	(.191)	-0.345	(.491)
β_4	1.168	(.180)	1.529	(.656)	1.174	(.249)	1.643	(.678)
ρ_{12}					0.433		0.335	
ρ_{23}					0.529		0.251	
ρ_{13}					0.351		0.228	
σ^2			1.900	(.736)			1.930	(.804)
β_1	-1.750	(.204)	-2.490	(.593)	-1.761	(.228)	-2.425	(.573)
β_2	-0.967	(.180)	-1.440	(.579)	-0.982	(.203)	-1.395	(.557)
β_3	-0.215	(.171)	-0.372	(.572)	-0.230	(.192)	-0.330	(.549)
β_4	1.227	(.182)	1.731	(.780)	1.235	(.251)	1.666	(.753)
ρ_{12}					0.386		0.118	
ρ_{23}					0.587		0.270	
ρ_{13}					0.351		0.135	
σ^2			2.713	(1.031)			2.446	(.984)

(continued)

	MODELS:							
Par.	Logistic		E B		G E E		Q E B	
β_1	-1.911	(.211)	-3.082	(.757)	-1.876	(.233)	-3.038	(.752)
β_2	-1.126	(.186)	-1.631	(.730)	-1.082	(.208)	-1.589	(.723)
β_3	-0.506	(.174)	-0.630	(.721)	-0.470	(.195)	-0.597	(.715)
β_4	1.469	(.186)	2.486	(1.007)	1.436	(.257)	2.429	(1.000)
ρ_{12}					0.515		0.0471	
ρ_{23}					0.573		0.186	
ρ_{13}					0.358		-0.0254	
σ^2			4.582	(1.586)			4.479	(1.573)

Table 4.13: Comparison of the QEB model with the logistic, empirical Bayes (EB), and GEE fits, assuming inclusion of a cluster-level covariate; (standard errors are indicated in brackets - each horizontal panel corresponds to one data set).

Examining the standard error of $\hat{\beta}_4$ for the GEE analyses, we observe that it seems to be inflated from that given by the logistic model, though not to the same extent as in the standard or quasi empirical Bayes models. Again, the variability in the parameter estimate of a cluster-level covariate is driven by the size of the random effects variance and not the intra-individual correlations, whereas that of the estimate of an individual-level covariate is largely determined by the intra-individual correlations, not the size of σ^2. Therefore the standard error for $\hat{\beta}_4$ given by GEE is approximately correct for the analyses presented in table 4.8, but not for those in table 4.13. Note however that no similar implication as set forth in the last paragraph holds for the GEE approach, with reference to individual-level covariates. In other words, even if interest is focussed only on individual-level covariates, GEE alone cannot be used to estimate $(\boldsymbol{\beta}, \boldsymbol{\rho})$ from model (4.4.1), since this approach estimates neither the correct (cluster-specific) coefficients $\boldsymbol{\beta}$, nor the appropriate conditional intra-individual correlations; see the discussion in the previous section. (On this final point consider for example the (albeit extreme) case of the fit of the

last data set in table 4.13, and note the large discrepancy in the correlation estimates).

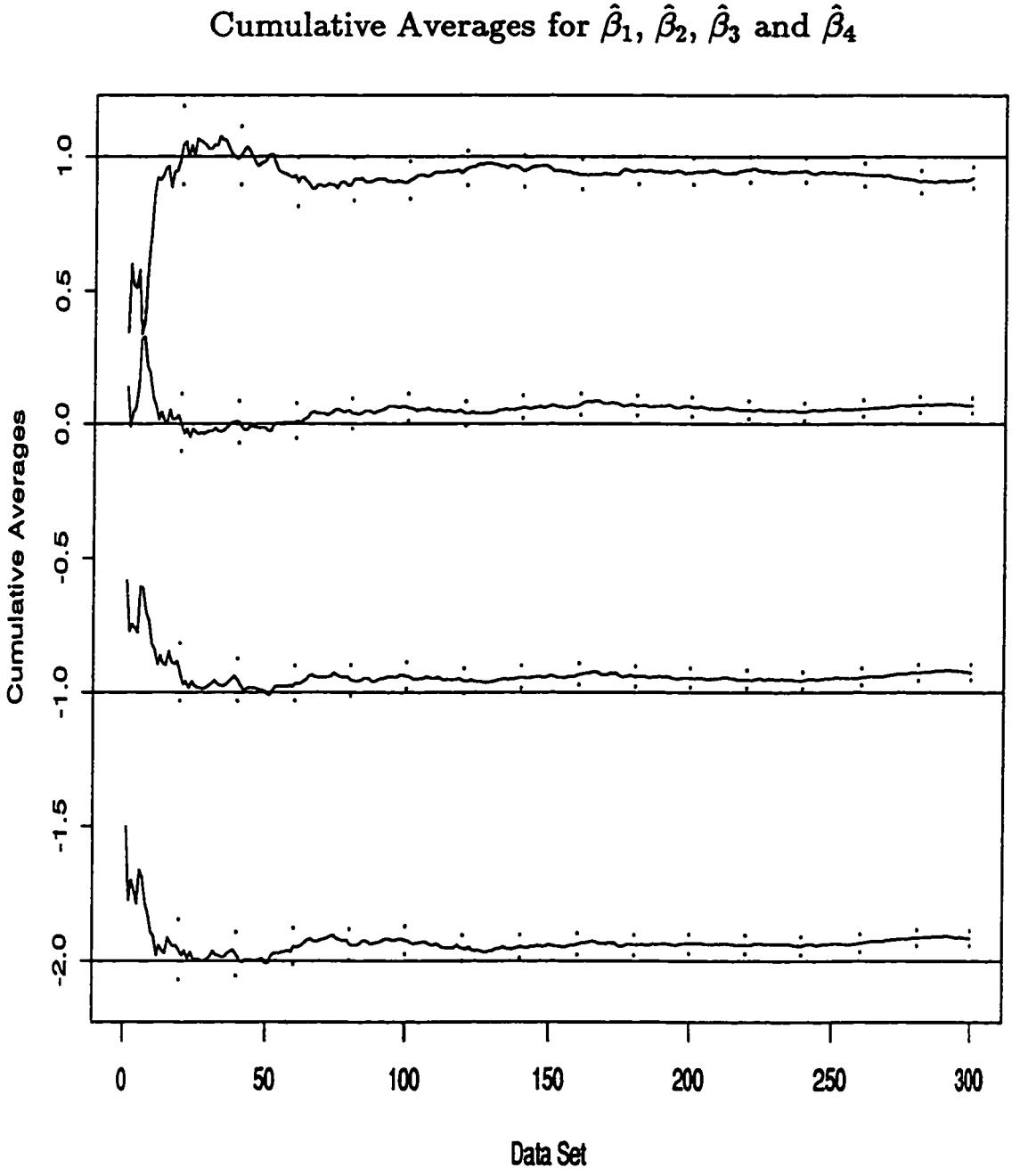


Figure 4.11: Cumulative averages of estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, assuming inclusion of a cluster-level covariate; points indicate ± 1 std. error.

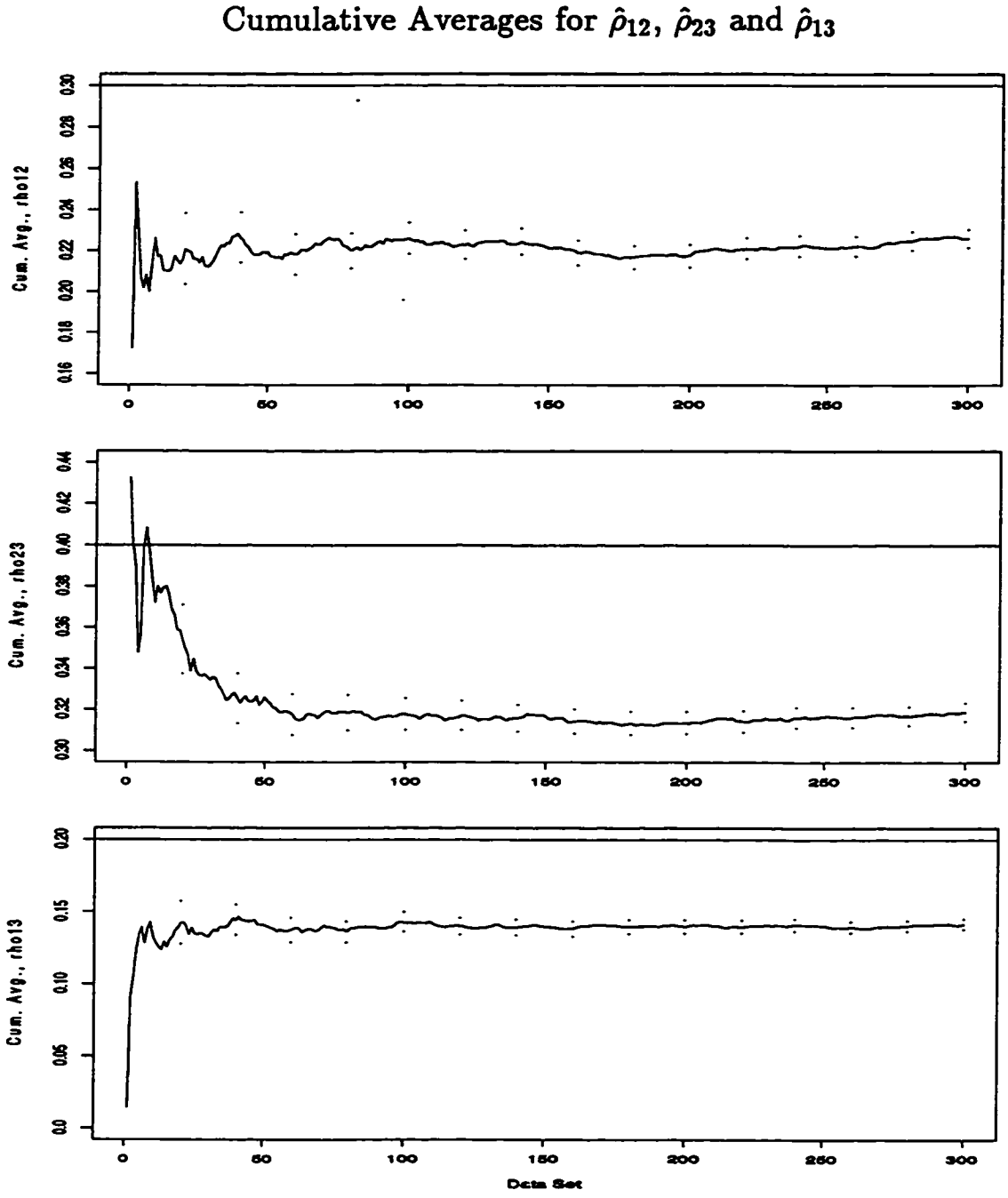


Figure 4.12: Cumulative averages of correlation estimates $\hat{\rho}_{12}$, $\hat{\rho}_{23}$, $\hat{\rho}_{13}$, assuming inclusion of a cluster-level covariate; points indicate ± 1 std. error.

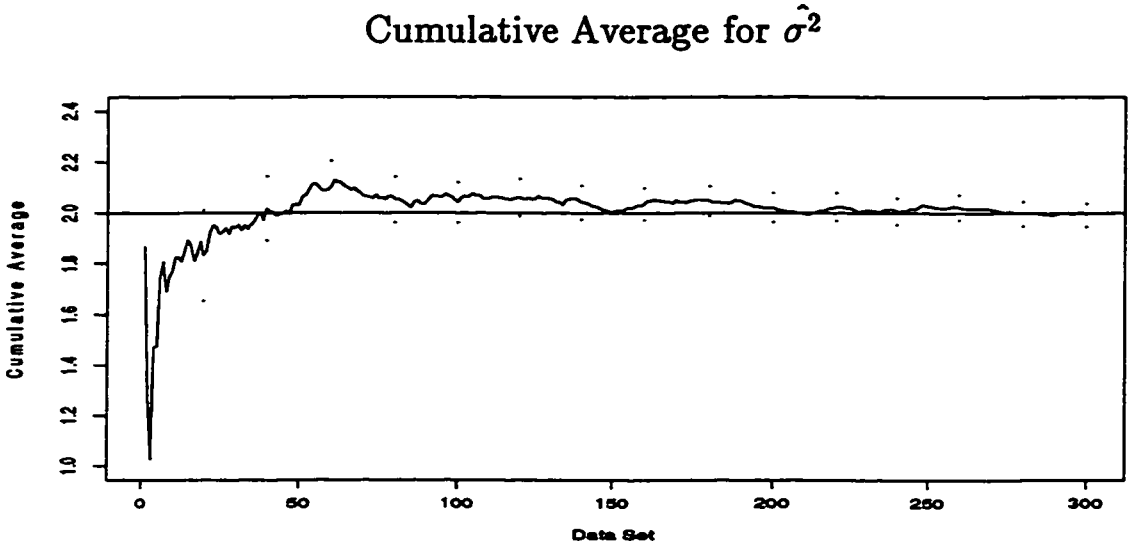


Figure 4.13: Cumulative average of variance estimates $\hat{\sigma^2}$, assuming inclusion of a cluster-level covariate; points indicate ± 1 std. error.

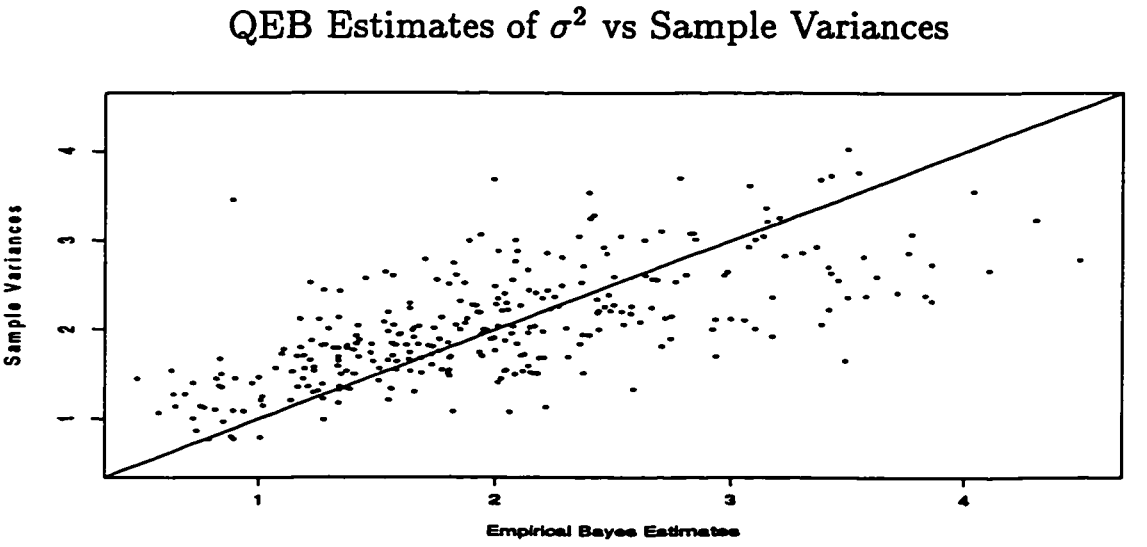


Figure 4.14: QEB Estimates of σ^2 versus sample variances of the (simulated) random effects, assuming inclusion of a cluster-level covariate.

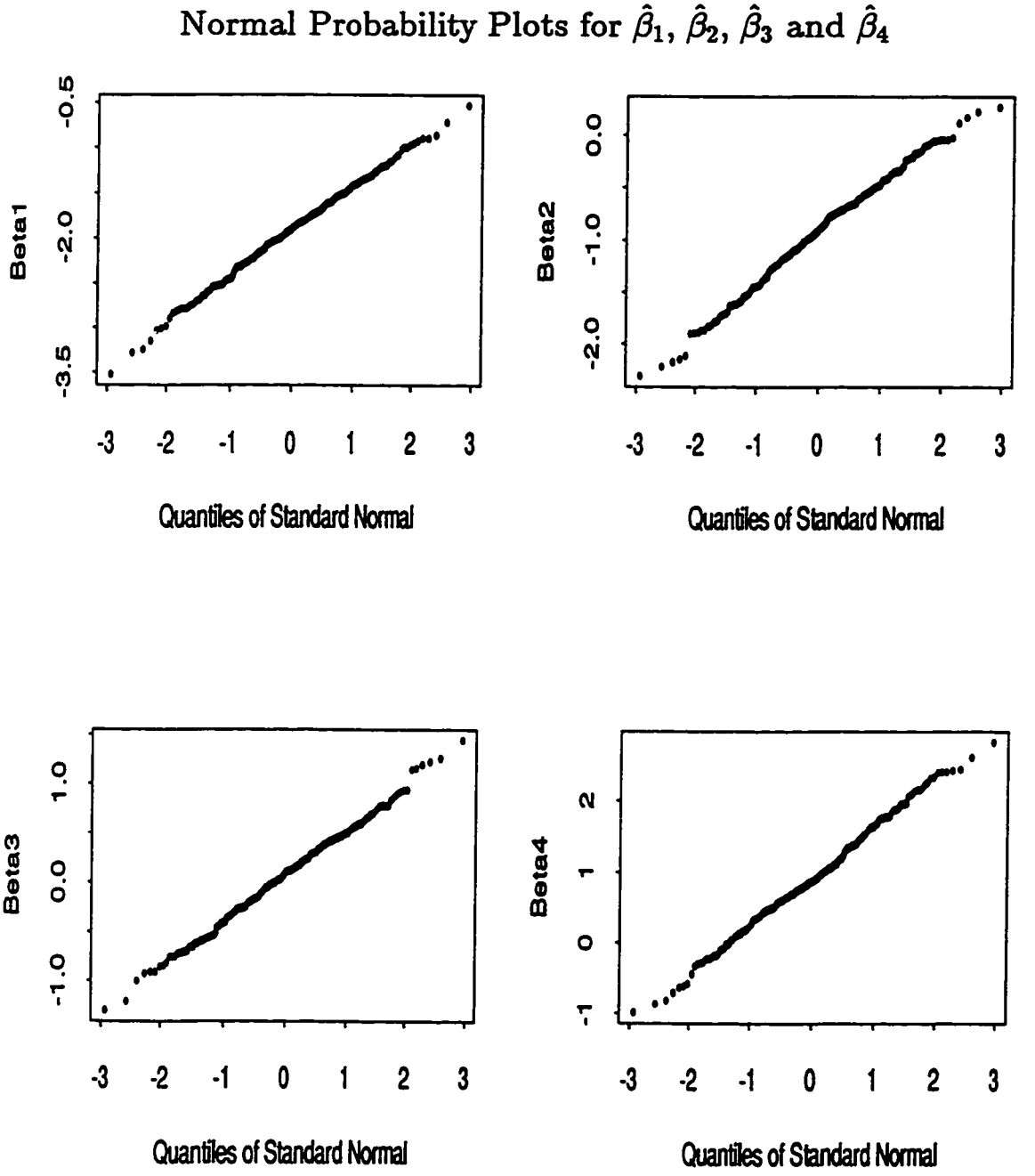


Figure 4.15: Normal probability plots for estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$, assuming inclusion of a cluster-level covariate.

4.6 Discussion

The composite quasi empirical Bayes model described in this chapter combines empirical Bayes methods and GEEs in a useful and relatively straightforward manner. It allows the modelling of more complicated correlation structures than either technique can reasonably support on its own, and can thus provide an analysis which better accomodates complex study or sampling designs such as that of the WSPP3.

We developed a robust covariance matrix for the fixed and random effects estimates from this model, adjusted for the fact that the prior variance σ^2, on which these estimates are conditioned, is not known but estimated empirically. This also allowed us to obtain an explicit estimate of the variability in $\hat{\sigma^2}$, which was previously not available. We hence described the simulation of data sets with a specified composite correlation stucture and various covariate patterns; these were then used to study the properties of the quasi empirical Bayes model.

It is interesting to note that somewhat better results, in terms of apparent unbiasedness of fixed effects estimates and confidence interval coverage properties, were observed for the data sets in section 4.5.2 which included an individual-level covariate, than in either sections 4.5.1 or 4.5.3. The reason for this phenomenon is unclear and deserves further study. At any rate, the bias in the fixed effects parameter estimates from the quasi empirical Bayes model does not appear to be severe. Neuhaus and Segal (1997) examine the performance of two approximate maximum likelihood estimators in the generalized linear mixed model framework. One is based on a second order Taylor series expansion of the integrated likelihood (3.6.3) about $\boldsymbol{b} = 0$, where $\boldsymbol{b}$ denotes the vector of random effects (see for example Longford (1994)), whereas the other is based on Breslow and Clayton's (1993) penalized quasi-likelihood (PQL) approach. Using this method the integrand for each cluster in (3.6.3) is approximated with a first order Taylor series expansion about the value $\hat{\boldsymbol{b}}_i$ which maximizes the integrand, and is then integrated analytically. Neuhaus and Segal's

findings suggest that the estimates obtained from these approaches are attenuated from the actual cluster-specific maximum likelihood estimates, more closely resembling population-averaged estimates. The authors show that in some cases this bias can be severe. For the PQL approach specifically, they show that parameter estimates will be close to population-averaged estimates if the estimated random effects variance is severely attenuated, which is common in PQL estimation. The quasi empirical Bayes approach advanced here is more similar to penalized quasi-likelihood than to the Taylor approximation method of Longford (1994). The empirical estimation of σ^2 however seems to provide only slightly biased estimates of the random effects variance, suggesting that the bias in the fixed effects parameter estimates should be less severe than one might expect from PQL. Of course one must bear in mind that quasi empirical Bayes estimation also adjusts for intra-individual correlation using a GEE approach, conditional on the cluster-level random effects. One should investigate further the impact on the bias of estimated covariate effects of this superimposed marginal method of accounting for the longitudinal dependence in the data. It would be of interest for instance to study model estimates as the degree of intra-individual correlation varies, given a fixed value of σ^2. Finally, we point out that the difference between population-averaged and cluster-specific parameter estimates is generally less dramatic than the examples in Neuhaus and Segal suggest. The authors discuss applications to matched pairs data and trachomal eye disease data, wherein clusters are of size 2 and the random effects variance in each case is very large. In studies such as the WSPP3, in which clusters contain many observations, the random effects variance is typically much smaller, and population-averaged and cluster-specific estimates tend to be more similar.

As a cautionary note, we must bear in mind that the results presented in section 4.5 were based on the assumption that students remain in the same school over time. In practice this assumption may be violated and we pointed out that an advantage of the quasi empirical Bayes model is its flexibility in allowing individuals to change schools over time. If a substantial number of such switches occur, or there is a systematic change

affecting all individuals at a given time point (such as the transition from elementary to highschools after grade 8), the results have to be interpreted more carefully in view of this.

Further discussion is deferred to Chapter 7, where we shall consider applications of the quasi empirical Bayes model in modelling the WSPP3 data.

4.7 Appendix

Here we describe the actual data simulation discussed in section 4.4. The following three subsections outline the generation of data with covariates for time only, with covariates for time plus an individual-level covariate, and with covariates for time plus a cluster-level covariate. Recall that the parameter values used were $(\beta_1, \beta_2, \beta_3, \beta_4) = (-2, -1, 0, 1)$, $(\rho_{12}, \rho_{23}, \rho_{13}) = (0.3, 0.4, 0.2)$, and $\sigma^2 = 2$.

A 1 Generating Data With Covariates for Time Only

The aim here is to generate data from model (4.4.1) where the only covariates are indicators for time. The marginal probabilities

$$p_{kj1} = \frac{e^{-2+b_k}}{1+e^{-2+b_k}}, \quad p_{kj2} = \frac{e^{-1+b_k}}{1+e^{-1+b_k}}, \quad p_{kj3} = \frac{e^{b_k}}{1+e^{b_k}}$$

must be recovered while ensuring that $\rho_{12} = 0.3$, $\rho_{23} = 0.4$ and $\rho_{13} = 0.2$.

Consider the first two time points. Begin by generating a sample of 20 independent observations from $N(0, \sigma^2)$, using these as the random effects $b_1, ..., b_{20}$. Noting that

$$Y_{kj1} \sim \text{Bin}(1, p_{kj1}), \quad j = 1, ..., 10, \quad k = 1, ..., 20,$$

generating the 200 observations for $t = 1$ is trivial. Correlation between observations on the same individual at times 1 and 2 is introduced by postulating a conditional probability

of response at time 2 given the outcome at time 1. Let

$$P(Y_{kj2}=1|Y_{kj1}=y_{kj1},b_k) \;=\; p_{kj}(\xi_k,\gamma_k,b_k) \;=\; \frac{e^{\xi_k+\gamma_k y_{kj1}+b_k}}{1+e^{\xi_k+\gamma_k y_{kj1}+b_k}}. \tag{4.7.1}$$

We require that

$$\rho_{12}=\mathrm{Corr}(Y_{kj1},Y_{kj2}|b_k)=\frac{\mathrm{E}(Y_{kj1}Y_{kj2}|b_k)-p_{kj1}p_{kj2}}{\sqrt{p_{kj1}(1-p_{kj1})p_{kj2}(1-p_{kj2})}}=0.3.$$

Now

$$\begin{aligned}\mathrm{E}(Y_{kj1}Y_{kj2}|b_k) &= P(Y_{kj1}=1,Y_{kj2}=1|b_k)\\ &= P(Y_{kj2}=1|Y_{kj1}=1,b_k)P(Y_{kj1}=1|b_k)\\ &= \frac{e^{\xi_k+\gamma_k+b_k}}{1+e^{\xi_k+\gamma_k+b_k}}\times\frac{e^{-2+b_k}}{1+e^{-2+b_k}}.\end{aligned}$$

Therefore ξ_k and γ_k must satisfy the following equation:

$$\frac{\dfrac{e^{\xi_k+\gamma_k+b_k}}{1+e^{\xi_k+\gamma_k+b_k}}\cdot\dfrac{e^{-2+b_k}}{1+e^{-2+b_k}}-\dfrac{e^{-2+b_k}}{1+e^{-2+b_k}}\cdot\dfrac{e^{-1+b_k}}{1+e^{-1+b_k}}}{\left(\dfrac{e^{-2+b_k}}{(1+e^{-2+b_k})^2}\cdot\dfrac{e^{-1+b_k}}{(1+e^{-1+b_k})^2}\right)^{1/2}}=0.3. \tag{4.7.2}$$

At the same time we must ensure $P(Y_{kj2}=1|b_k)=p_{kj2}=e^{-1+b_k}/(1+e^{-1+b_k})$; observe that

$$\begin{aligned}P(Y_{kj2}=1|b_k) &= P(Y_{kj2}=1|Y_{kj1}=1,b_k)P(Y_{kj1}=1|b_k)+\\ &\quad P(Y_{kj2}=1|Y_{kj1}=0,b_k)P(Y_{kj1}=0|b_k).\end{aligned}$$

Hence ξ_k and γ_k must also satisfy the equation

$$\frac{e^{\xi_k+\gamma_k+b_k}}{1+e^{\xi_k+\gamma_k+b_k}} \cdot \frac{e^{-2+b_k}}{1+e^{-2+b_k}} + \frac{e^{\xi_k+b_k}}{1+e^{\xi_k+b_k}} \cdot \frac{1}{1+e^{-2+b_k}} = \frac{e^{-1+b_k}}{1+e^{-1+b_k}}. \tag{4.7.3}$$

Therefore, for each random effect b_k (each school) we obtain two equations in two unknowns which can be readily solved for ξ_k and γ_k. Having obtained these values, equation (4.7.1) can be used together with the data from time 1 to generate the time 2 observations:

$$Y_{kj2}|Y_{kj1} \sim \text{Bin}(1, p_{kj}(\xi_k, \gamma_k, b_k)), \quad j = 1, ..., 10, \quad k = 1, ..., 20.$$

The Y_{kj2} generated from this conditional model have the desired marginal mean and are appropriately correlated with Y_{kj1}.

In a similar fashion, correlation between observations on the same individual at times 2 and 3 and at times 1 and 3 is introduced by postulating a conditional probability of response at time 3, given the outcomes at times 1 and 2. Let

$$\begin{aligned} P(Y_{kj3} = 1|Y_{kj2} = y_{kj2}, Y_{kj1} = y_{kj1}, b_k) &= p_{kj}(\zeta_k, \vartheta_{2k}, \vartheta_{1k}, b_k) \\ &= \frac{e^{\zeta_k+\vartheta_{2k}y_{kj2}+\vartheta_{1k}y_{kj1}+b_k}}{1+e^{\zeta_k+\vartheta_{2k}y_{kj2}+\vartheta_{1k}y_{kj1}+b_k}}. \end{aligned} \tag{4.7.4}$$

We require that

$$\begin{aligned} \rho_{23} &= \text{Corr}(Y_{kj2}, Y_{kj3}|b_k) = \frac{\text{E}(Y_{kj2}Y_{kj3}|b_k) - p_{kj2}p_{kj3}}{\sqrt{p_{kj2}(1-p_{kj2})p_{kj3}(1-p_{kj3})}} = 0.4 \quad \text{and} \\ \rho_{13} &= \text{Corr}(Y_{kj1}, Y_{kj3}|b_k) = \frac{\text{E}(Y_{kj1}Y_{kj3}|b_k) - p_{kj1}p_{kj3}}{\sqrt{p_{kj1}(1-p_{kj1})p_{kj3}(1-p_{kj3})}} = 0.2. \end{aligned}$$

Consider first ρ_{23}.

$$\text{E}(Y_{kj2}Y_{kj3}|b_k) = \sum_{y_{kj1}=0}^{1} P(Y_{kj3} = 1|Y_{kj2} = 1, Y_{kj1} = y_{kj1}, b_k) \times$$

$$P(Y_{kj2} = 1|Y_{kj1} = y_{kj1}, b_k)P(Y_{kj1} = y_{kj1}|b_k).$$

Abbreviating (4.7.4) by $P(Y_{kj3} = 1|y_{kj2}, y_{kj1}, b_k)$ and (4.7.1) by $P(Y_{kj2} = 1|y_{kj1}, b_k)$, it follows that ζ_k, ϑ_{2k} and ϑ_{1k} must satisfy the equation

$$\begin{aligned}\{[P(Y_{kj3} = 1|1, 0, b_k) \cdot P(Y_{kj2} = 1|0, b_k) \cdot (1 - p_{kj1}) + \\ P(Y_{kj3} = 1|1, 1, b_k) \cdot P(Y_{kj2} = 1|1, b_k) \cdot p_{kj1}] - p_{kj2}p_{kj3}\} \times \\ \{p_{kj2}(1 - p_{kj2})p_{kj3}(1 - p_{kj3})\}^{-1/2} = 0.4. \qquad (4.7.5)\end{aligned}$$

Considering now ρ_{13}, we have

$$\begin{aligned}\mathrm{E}(Y_{kj1}Y_{kj3}|b_k) = \sum_{y_{kj2}=0}^{1} P(Y_{kj3} = 1|Y_{kj2} = y_{kj2}, Y_{kj1} = 1, b_k) \times \\ P(Y_{kj2} = y_{kj2}|Y_{kj1} = 1, b_k)P(Y_{kj1} = 1|b_k).\end{aligned}$$

Therefore ζ_k, ϑ_{2k} and ϑ_{1k} must also satisfy the equation

$$\begin{aligned}\{[P(Y_{kj3} = 1|0, 1, b_k) \cdot P(Y_{kj2} = 0|1, b_k) \cdot p_{kj1} + \\ P(Y_{kj3} = 1|1, 1, b_k) \cdot P(Y_{kj2} = 1|1, b_k) \cdot p_{kj1}] - p_{kj1}p_{kj3}\} \times \\ \{p_{kj1}(1 - p_{kj1})p_{kj3}(1 - p_{kj3})\}^{-1/2} = 0.2. \qquad (4.7.6)\end{aligned}$$

Finally, we must ensure that $P(Y_{kj3} = 1|b_k) = p_{kj3} = e^{b_k}/(1 + e^{b_k})$. Noting that

$$\begin{aligned}P(Y_{kj3} = 1|b_k) = \sum_{y_{kj1}=0}^{1} \sum_{y_{kj2}=0}^{1} P(Y_{kj3} = 1|y_{kj2}, y_{kj1}, b_k) \times \\ P(Y_{kj2} = y_{kj2}|y_{kj1}, b_k)P(Y_{kj1} = y_{kj1}|b_k),\end{aligned}$$

we see that ζ_k, ϑ_{2k} and ϑ_{1k} also have to satisfy

$$\begin{aligned} P(Y_{kj3}=1|0,0,b_k)\cdot P(Y_{kj2}=0|0,b_k)\cdot(1-p_{kj1}) &+ \\ P(Y_{kj3}=1|1,0,b_k)\cdot P(Y_{kj2}=1|0,b_k)\cdot(1-p_{kj1}) &+ \\ P(Y_{kj3}=1|0,1,b_k)\cdot P(Y_{kj2}=0|1,b_k)\cdot p_{kj1} &+ \\ P(Y_{kj3}=1|1,1,b_k)\cdot P(Y_{kj2}=1|1,b_k)\cdot p_{kj1} &= \frac{e^{b_k}}{1+e^{b_k}}. \end{aligned} \tag{4.7.7}$$

Hence at this stage we obtain for each random effect b_k three equations in three unknowns ((4.7.5), (4.7.6) and (4.7.7)) which can be solved for ζ_k, ϑ_{2k} and ϑ_{1k}. Having obtained these values, equation (4.7.4) can be used together with the data from time points 2 and 1 to generate the time 3 observations:

$$Y_{kj3}|Y_{kj2},Y_{kj1} \sim \text{Bin}(1,p_{kj}(\zeta_k,\vartheta_{2k},\vartheta_{1k},b_k)), \quad j=1,...,10, \quad k=1,...,20.$$

Similar to the previous result, the Y_{kj3} generated from this conditional model have the desired marginal probability and are appropriately correlated with both Y_{kj1} and Y_{kj2}.

In summary, we obtain the following algorithm for generating a data set with covariates for time only, exhibiting the correlation structure described here:

1. Generate a sample of 20 independent observations from $N(0,\sigma^2)$, using these as the random effects $b_1,...,b_{20}$.
2. Generate Y_{kj1} from $\text{Bin}(1,p_{kj1})$, $j=1,...,10$, $k=1,...,20$.
3. For each b_k, $k=1,...,20$, solve equations (4.7.2) and (4.7.3) for ξ_k and γ_k.
4. Generate Y_{kj2} from $\text{Bin}(1,p_{kj}(\xi_k,\gamma_k,b_k))$, $j=1,...,10$, $k=1,...,20$.
5. For each b_k, $k=1,...,20$, solve equations (4.7.5), (4.7.6) and (4.7.7) for ζ_k, ϑ_{2k} and ϑ_{1k}.

6. Generate Y_{kj3} from $\text{Bin}(1, p_{kj}(\zeta_k, \vartheta_{2k}, \vartheta_{1k}, b_k))$, $j = 1, ..., 10$, $k = 1, ..., 20$.

A 2 Generating Data With an Individual-Level Covariate

The method described in the previous section generalizes in a straightforward manner to the problem of simulating data from model (4.4.1), where the covariate structure includes indicators for time as well as an individual-level covariate. Suppose as we have above that this additional covariate is an indicator for gender, with half the students in each school being males and half females. In general the marginal probabilities

$$p_{kjt}(g_{kj}) = P(Y_{kjt} = 1 | g_{kj}, b_k) = \frac{e^{\beta_t + \beta_4 g_{kj} + b_k}}{1 + e^{\beta_t + \beta_4 g_{kj} + b_k}}$$

must be recovered, which we can categorize as either $p_{kjt}(1)$ or $p_{kjt}(0)$, denoting probabilities associated with females and males, respectively. Therefore this yields six distinct marginal probabilities for each school:

$$p_{kj1}(1) = \frac{e^{-2+1+b_k}}{1 + e^{-2+1+b_k}}, \quad p_{kj2}(1) = \frac{e^{-1+1+b_k}}{1 + e^{-1+1+b_k}}, \quad p_{kj3}(1) = \frac{e^{1+b_k}}{1 + e^{1+b_k}}$$
$$p_{kj1}(0) = \frac{e^{-2+b_k}}{1 + e^{-2+b_k}}, \quad p_{kj2}(0) = \frac{e^{-1+b_k}}{1 + e^{-1+b_k}}, \quad p_{kj3}(0) = \frac{e^{b_k}}{1 + e^{b_k}}.$$

Consider again the first two time points. Begin by obtaining a sample of 20 independent random effects from $N(0, \sigma^2)$ and generate the observations for $t = 1$ according to

$$Y_{kj1} \sim \text{Bin}(1, p_{kj1}(g_{kj})), \quad j = 1, ..., 10, \quad k = 1, ..., 20.$$

Once again, correlation between observations on the same individual at times 1 and 2 is introduced by specifying a conditional probability of response at time 2 given the outcome at time 1. However to ensure that the intra-individual correlations are the same for males

and females, we must specify separate probabilities for each gender. Let

$$\begin{aligned} \text{logit}\, P(Y_{kj2} = 1 | y_{kj1}, g_{kj}, b_k) &= \text{logit}\ p_{kj}(\xi_k, \gamma_k, g_{kj}, b_k) \\ &= \begin{cases} \xi_{kF} + \gamma_{kF} y_{kj1} + b_k & \text{if } g_{kj} = 1 \\ \xi_{kM} + \gamma_{kM} y_{kj1} + b_k & \text{if } g_{kj} = 0. \end{cases} \end{aligned} \tag{4.7.8}$$

Now consider the case for females. We require that

$$\rho_{12} = \text{Corr}(Y_{kj1}, Y_{kj2} | g_{kj} = 1, b_k) = \frac{\text{E}(Y_{kj1} Y_{kj2} | g_{kj} = 1, b_k) - p_{kj1}(1) p_{kj2}(1)}{\sqrt{p_{kj1}(1)(1 - p_{kj1}(1)) p_{kj2}(1)(1 - p_{kj2}(1))}} = 0.3.$$

Therefore ξ_{kF} and γ_{kF} must satisfy the equation

$$\frac{\dfrac{e^{\xi_{kF}+\gamma_{kF}+b_k}}{1+e^{\xi_{kF}+\gamma_{kF}+b_k}} \cdot \dfrac{e^{-2+1+b_k}}{1+e^{-2+1+b_k}} - \dfrac{e^{-2+1+b_k}}{1+e^{-2+1+b_k}} \cdot \dfrac{e^{-1+1+b_k}}{1+e^{-1+1+b_k}}}{\left(\dfrac{e^{-2+1+b_k}}{(1+e^{-2+1+b_k})^2} \cdot \dfrac{e^{-1+1+b_k}}{(1+e^{-1+1+b_k})^2} \right)^{1/2}} = 0.3. \tag{4.7.9}$$

At the same time we must ensure that $P(Y_{kj2} = 1 | g_{kj} = 1, b_k) = p_{kj2}(1) = e^{-1+1+b_k}/(1 + e^{-1+1+b_k})$; this requires that ξ_{kF} and γ_{kF} must also satisfy

$$\begin{aligned} \frac{e^{\xi_{kF}+\gamma_{kF}+b_k}}{1+e^{\xi_{kF}+\gamma_{kF}+b_k}} \cdot \frac{e^{-2+1+b_k}}{1+e^{-2+1+b_k}} + \frac{e^{\xi_{kF}+b_k}}{1+e^{\xi_{kF}+b_k}} \cdot \frac{1}{1+e^{-2+1+b_k}} \\ = \frac{e^{-1+1+b_k}}{1+e^{-1+1+b_k}}. \end{aligned} \tag{4.7.10}$$

Therefore, for each random effect b_k we again obtain two equations in two unknowns which can be solved for ξ_{kF} and γ_{kF}. The development for males proceeds in exactly the same manner. From the conditions

$$\rho_{12} = \text{Corr}(Y_{kj1}, Y_{kj2} | g_{kj} = 0, b_k) = 0.3$$

and

$$P(Y_{kj2}=1|g_{kj}=0,b_k) \;=\; p_{kj2}(0) \;=\; \frac{e^{-1+b_k}}{1+e^{-1+b_k}}$$

we obtain two equations similar to (4.7.9) and (4.7.10); simply replace ξ_{kF} and γ_{kF} with ξ_{kM} and γ_{kM}, and the marginal probabilities $p_{kjt}(1)$ with $p_{kjt}(0)$. These equations can then be solved for ξ_{kM} and γ_{kM}. Having obtained the four values ξ_{kF}, γ_{kF}, ξ_{kM} and γ_{kM}, equation (4.7.8) can be used together with the data from time 1 to generate the time 2 observations:

$$Y_{kj2}|Y_{kj1} \sim \text{Bin}(1, p_{kj}(\xi_k, \gamma_k, g_{kj}, b_k)), \quad j=1,...,10, \;\; k=1,...,20.$$

Analogous to (4.7.4) we next postulate gender-specific probabilities of response at time 3, given the outcomes at times 1 and 2:

$$\begin{aligned}\text{logit}\, P(Y_{kj3}=1|y_{kj2},y_{kj1},g_{kj},b_k) \;&=\; \text{logit}\; p_{kj}(\zeta_k,\vartheta_{2k},\vartheta_{1k},g_{kj},b_k)\\ &=\; \begin{cases}\zeta_{kF}+\vartheta_{2kF}y_{kj2}+\vartheta_{1kF}y_{kj1}+b_k & \text{if } g_{kj}=1\\ \zeta_{kM}+\vartheta_{2kM}y_{kj2}+\vartheta_{1kM}y_{kj1}+b_k & \text{if } g_{kj}=0.\end{cases}\end{aligned} \tag{4.7.11}$$

Considering again the case for females, the conditions

$$\begin{aligned}\rho_{23} \;&=\; \text{Corr}(Y_{kj2},Y_{kj3}|g_{kj}=1,b_k)=0.4,\\ \rho_{13} \;&=\; \text{Corr}(Y_{kj1},Y_{kj3}|g_{kj}=1,b_k)=0.2 \quad \text{and}\\ p_{kj3}(1) \;&=\; P(Y_{kj3}=1|g_{kj}=1,b_k) \;=\; \frac{e^{1+b_k}}{1+e^{1+b_k}}\end{aligned}$$

lead to the equations

$$\begin{aligned}\{[P(Y_{kj3}=1|1,0,g_{kj}=1,b_k)\cdot P(Y_{kj2}=1|0,g_{kj}=1,b_k)\cdot(1-p_{kj1}(1)) \;\;+\\ P(Y_{kj3}=1|1,1,g_{kj}=1,b_k)\cdot P(Y_{kj2}=1|1,g_{kj}=1,b_k)\cdot p_{kj1}(1)] \;\;-\\ p_{kj2}(1)p_{kj3}(1)\}\times\{p_{kj2}(1)(1-p_{kj2}(1))p_{kj3}(1)(1-p_{kj3}(1))\}^{-1/2} \;=\; 0.4,\end{aligned} \tag{4.7.12}$$

$$\begin{aligned}\{[P(Y_{kj3}=1|0,1,g_{kj}=1,b_k)\cdot P(Y_{kj2}=0|1,g_{kj}=1,b_k)\cdot p_{kj1}(1) \quad &+\\ P(Y_{kj3}=1|1,1,g_{kj}=1,b_k)\cdot P(Y_{kj2}=1|1,g_{kj}=1,b_k)\cdot p_{kj1}(1)] \quad &-\\ p_{kj1}(1)p_{kj3}(1)\}\times\{p_{kj1}(1)(1-p_{kj1}(1))p_{kj3}(1)(1-p_{kj3}(1))\}^{-1/2} \quad &= \quad 0.2 \qquad (4.7.13)\end{aligned}$$

and

$$\begin{aligned}P(Y_{kj3}=1|0,0,g_{kj}=1,b_k)\cdot P(Y_{kj2}=0|0,g_{kj}=1,b_k)\cdot(1-p_{kj1}(1)) \quad &+\\ P(Y_{kj3}=1|1,0,g_{kj}=1,b_k)\cdot P(Y_{kj2}=1|0,g_{kj}=1,b_k)\cdot(1-p_{kj1}(1)) \quad &+\\ P(Y_{kj3}=1|0,1,g_{kj}=1,b_k)\cdot P(Y_{kj2}=0|1,g_{kj}=1,b_k)\cdot p_{kj1}(1) \quad &+\\ P(Y_{kj3}=1|1,1,g_{kj}=1,b_k)\cdot P(Y_{kj2}=1|1,g_{kj}=1,b_k)\cdot p_{kj1}(1) &\\ = \frac{e^{1+b_k}}{1+e^{1+b_k}}, \qquad (4.7.14)\end{aligned}$$

which can be solved for ζ_{kF}, ϑ_{2kF} and ϑ_{1kF}. Again, the analogous equations for ζ_{kM}, ϑ_{2kM} and ϑ_{1kM} are obtained in the same way. Simply replace all occurrences of ζ_{kF}, ϑ_{2kF} and ϑ_{1kF} in (4.7.12), (4.7.13) and (4.7.14) with ζ_{kM}, ϑ_{2kM} and ϑ_{1kM}, and the marginal probabilities $p_{kjt}(1)$ with $p_{kjt}(0)$. Having obtained the six values ζ_{kF}, ϑ_{2kF}, ϑ_{1kF}, ζ_{kM}, ϑ_{2kM} and ϑ_{1kM}, equation (4.7.11) is used together with the data from time points 2 and 1 to generate the time 3 observations:

$$Y_{kj3}|Y_{kj2},Y_{kj1}\sim \text{Bin}(1,p_{kj}(\zeta_k,\vartheta_{2k},\vartheta_{1k},g_{kj},b_k)), \quad j=1,...,10, \quad k=1,...,20.$$

Hence we obtain the following algorithm for generating a data set with indicators for time, as well as an individual-level covariate:

1. Generate a sample of 20 independent observations from $N(0,\sigma^2)$, using these as the random effects $b_1,...,b_{20}$.

2. Generate Y_{kj1} from $\text{Bin}(1, p_{kj1}(g_{kj}))$, $j = 1, ..., 10$, $k = 1, ..., 20$.

3. For each b_k, $k = 1, ..., 20$, solve equations (4.7.9) and (4.7.10) for ξ_{kF} and γ_{kF}, and the analogous equations for ξ_{kM} and γ_{kM}.

4. Generate Y_{kj2} from $\text{Bin}(1, p_{kj}(\xi_k, \gamma_k, g_{kj}, b_k))$, $j = 1, ..., 10$, $k = 1, ..., 20$.

5. For each b_k, $k = 1, ..., 20$, solve equations (4.7.12), (4.7.13) and (4.7.14) for ζ_{kF}, ϑ_{2kF} and ϑ_{1kF}, and the analogous equations for ζ_{kM}, ϑ_{2kM} and ϑ_{1kM}.

6. Generate Y_{kj3} from $\text{Bin}(1, p_{kj}(\zeta_k, \vartheta_{2k}, \vartheta_{1k}, g_{kj}, b_k))$, $j = 1, ..., 10$, $k = 1, ..., 20$.

A 3 Generating Data With a Cluster-Level Covariate

One can simulate data from model (4.4.1) where the covariate structure includes indicators for time as well as a cluster-level covariate, using the same equations as developed in the last section. We simply bear in mind that this additional covariate is now on the level of the school, so that we can express the required marginal probabilities as

$$p_{kjt}(g_k) = P(Y_{kjt} = 1 | g_k, b_k) = \frac{e^{\beta_t + \beta_4 g_k + b_k}}{1 + e^{\beta_t + \beta_4 g_k + b_k}},$$

which we shall categorize as either $p_{kjt}(1)$ or $p_{kjt}(0)$, denoting probabilities associated with, say, treatment and control schools respectively. Therefore this yields six distinct marginal probabilities (now three for each type of school), identical to the ones given in the previous section. From here the same arguments apply as before, but instead of carrying calculations through for both gender types in each school, we perform one set of calculations for each of the two types of schools. One can then apply an algorithm very similar to that described at the end of section A 2, the only notable difference being, again, the fact that in steps (3) and (5) the two systems of two equations and two systems of three equations respectively

are solved for the two different types of schools, rather than for the two gender types within each school.

Chapter 5

Approximating Correlation Structures in Clustered Binary Data Using Random Effects Models

5.1 Introduction

In Chapter 4 we described a general method of addressing the composite correlation structure inherent in longitudinal data which are also cross-sectionally grouped. In this chapter we focus specifically and in greater detail on the element of cross-sectional clustering and suggest applications of random effects models which can be interpreted as modelling both the mean as well as the correlation structure of the data. We will show that such models can lead to a significant improvement in fit over other simpler random effects models, by making the most effective use of covariates thought to be related to the correlation structure (concerning model fit, see in particular Chapter 6). Such covariates might well exist, and may or may not be associated with the mean specification of the model; numerous examples will be discussed throughout the chapter - we shall begin with a simple motivating

illustration.

Consider data from WSPP1, the first study (1979-1982) in the series of Waterloo Smoking Prevention Projects. This was a relatively small-scale formative evaluation of a school-based smoking prevention program. Twenty-two schools from two school boards were randomly assigned to either a treatment or a control condition. The core WSPP program was delivered to students in schools within the treatment group in grade 6, with booster sessions given in each of grades 7 and 8. The table below lists the data on the students in grade 8, where r is the number of smokers out of n students observed at a given school. We use the abbreviations "WSSB" to stand for Waterloo Separate School Board and "OPSB" for Oxford Public School Board.

Control Schools						Treatment Schools					
WSSB			OPSB			WSSB			OPSB		
r	n	r/n	r	n	r/n	r	n	r/n	r	n	r/n
9	19	0.47	7	19	0.37	6	23	0.26	3	18	0.17
1	17	0.06	5	13	0.38	2	29	0.07	11	28	0.39
9	24	0.38	5	18	0.28	0	16	0.00	4	30	0.13
18	32	0.56	4	18	0.22	8	22	0.36	0	17	0.00
3	11	0.27	0	8	0.00	7	21	0.33	1	22	0.05
5	12	0.42				14	22	0.64			

The fit of a standard logistic model to these data, expressing the response proportion in each school as a function of treatment condition and school board (the interaction between the two factors was found to be insignificant) produces a deviance of 71.82 on 19 degrees of freedom, showing clear evidence of overdispersion; see the first panel of table 5.1. A logical next step might be to assume that the unexplained school-to-school heterogeneity, having adjusted for treatment and school board effects, is due to certain unobserved effects deriving from the same distribution. This could be modelled by the simple logistic-normal

	Logistic Model		Simple R. E. Model		Refined R. E. Model	
	est.	s.e.	est.	s.e.	est.	s.e.
β_0	-1.0105	(.2120)	-1.2149	(.3990)	-0.9391	(.2604)
β_1	-0.5648	(.2169)	-0.6433	(.4338)	-1.1994	(.2884)
β_2	0.5949	(.2250)	0.6554	(.4381)	0.7329	(.2917)
γ					2.2159	(.9515)
α			-0.3859	(.5220)	-2.7092	(1.633)
(σ^2)			0.6799		0.06659	
llik	-252.00		-241.97		-236.70	
D (d.f.)	71.82 (19)		51.76 (18)		41.22 (17)	

Table 5.1: Motivating example: various model fits to the WSPP1 data; (last two lines report the log-likelihood and model deviance, respectively).

model

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 trt_i + \beta_2 brd_i + b_i, \qquad b_i \overset{iid}{\sim} N(0, \sigma_1^2),$$

where p_i represents the proportion of smokers in school i, trt_i takes value 1 if school i is in the treatment condition and 0 otherwise, and similarly brd_i takes value 1 if school i is in the Waterloo Separate School Board, and 0 otherwise. The fit of this model is summarized in the second panel of 5.1. Maximum likelihood estimates were computed, using numerical integration to obtain the marginal likelihood; to achieve greater numerical stability we parameterized σ_1^2 as e^{α_1} and maximized the likelihood in $(\beta_0, \beta_1, \beta_2, \alpha_1)$. A clear improvement in fit is realized over the logistic model; at the same time, whereas the effect of both treatment condition and school board seemed to be significant in the logistic model, neither of these factors retains statistical significance under the simple random effects model above.

In trying to describe school-to-school variability more carefully one will notice a striking feature in this data set: in each group of schools defined by a common treatment - school board combination, one school clearly stands out by reporting a proportion of smokers

much larger or smaller than the remaining schools in the same group. Among the control schools, the observed proportions 1/17 and 0/8 in the WSSB and OPSB respectively seem unusually small, and among the treatment schools the rates 14/22 and 11/28 in the same two boards are strikingly large. If in fact there were some justification beyond inspection of the data for treating these schools as different from the rest, one could entertain the following refined random effects model formulation:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 trt_i + \beta_2 brd_i + e^{\gamma z_i} b_i, \qquad b_i \overset{iid}{\sim} N(0, \sigma_2^2),$$

where z_i is an indicator taking value 1 for the four schools in question and 0 otherwise. We surmise that most of the unexplained variability in the logistic model, captured as best as possible by the estimate of σ_1^2 in the simple random effects model, is in fact due to the widely varying responses in only these four schools. Hence when fitting the above model we should expect a rather large point estimate $\hat{\gamma}$, and in contrast $\hat{\sigma}_2^2$ to be quite small in comparison to $\hat{\sigma}_1^2$. Referring to the last panel in 5.1 we see that this is indeed the case. The impact of the random effects b_i is very much smaller for the 18 schools not highlighted, than estimated under the first random effects model. (Note the difference by a factor of 10 in $\hat{\sigma}_2^2$ vs $\hat{\sigma}_1^2$). On the other hand this impact is greatly inflated (by an estimated factor of $e^{2.2159} = 9.2$) for the four schools in question. The refined model achieves an additional significant increase in likelihood over the simpler random effects model. It is also interesting to note that while both these models produce standard errors for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that are larger than those from the logistic model, the refined model, which better captures the covariance structure in the data, has smaller standard errors than the simpler model. This need not be the case necessarily, but is not inconsistent with Neuhaus et al. (1992) who examine the effects of mixture distribution misspecification in mixed effects models; further discussion is provided in sections 5.5 and 5.6. In particular one will note that in describing the nature of the overdispersion as we have in the refined random effects model, the impact

of both treatment condition and school board are again judged to be significant. Beyond that the magnitude of the coefficient for treatment condition is twice as large in the refined model as compared to the logistic, which is perhaps not surprising since the refined model attenuates the impact of the two treatment schools with the highest smoking rates, and the two control schools with the lowest, suggesting a stronger than actual treatment effect. (Grouping the two treatment schools with the lowest smoking rates together with the four previously mentioned schools and refitting the refined random effects model produced a more moderate estimate for treatment condition).

When actually analyzing data we certainly do not recommend approaching model selection in the above manner. The above example was simply meant to illustrate the fact that there may exist situations in which the nature of the apparent overdispersion is such that it can be described by a functional relationship, quantifiable in terms of one (or several) covariates, in much the same way as the mean is specified. Further, a better model fit may be achieved by specifying such a relationship, and inferences might also be affected. The random effects models we consider in this chapter are all of the same structure as the refined model presented here, and are described and interpreted in sections 5.2 and 5.3. In particular, in section 5.3 we interpret such models as equivalent to particular realizations of population-averaged representations, with certain intra-cluster correlation structures. Through these models it is possible to meet the main objective in the modelling process, namely to obtain "good" fixed effects estimates (in the sense of small bias and asymptotically valid standard errors), as well as a reasonable reflection of the correlation structure in the data.

The approach described here allows for greater flexibility in allowing for varying intra-cluster correlation than does GEE, mainly because it can incorporate covariate information in a natural way, whereas GEE cannot. On the other hand, intra-cluster correlation is modelled in a more explicit manner through GEE than through random effects models. However a major advantage of the random effects formulation is the fact that it admits

a likelihood-based analysis, thus facilitating the choice between competing nested models. In section 5.4 we discuss some issues relating to the choice of particular random effects models, and in section 5.5 we examine the results of simulations performed to investigate questions of power and the effects of model misspecification.

Note that the methods described here could also be used to extend, in a fairly straightforward manner, the quasi empirical Bayes model described in the previous chapter. This however is not considered further at this point.

5.2 A General Class of Random Effects Models

5.2.1 The Model

For the sake of consistency with section 3.6, and to emphasize the shift in focus from the composite modelling approach in the last chapter to issues related specifically to cross-sectional clustering, we again adopt the notation used in section 3.6, where we initially discussed random effects models. For the analysis of binary data, assuming the logit link, we consider models of the following form for the probability p_{ij} that the jth individual in cluster i has response $Y_{ij} = 1$:

$$\log \frac{p_{ij}}{1-p_{ij}} = x_{ij}'\beta + f(z_{ij};\gamma)\cdot b_i, \qquad b_i \sim N(0,\sigma^2). \tag{5.2.1}$$

Assume as usual that two observations from the same cluster are independent only conditional on the random effect for that cluster; i.e., that they are marginally correlated. Naturally one could write down a similar formulation for other link functions or data types; for dichotomous data in particular, (5.2.1) offers a convenient means of addressing both variation in the mean response, as well as allowing for an indirect modelling of the correlation structure through the function $f(z_{ij};\gamma)$. In multiplying the cluster-specific

random effect, this function, which can be specific to the individual, either inflates or attenuates the random effect, thus tailoring its impact for each specific cluster, perhaps even each individual. The term $f(z_{ij};\gamma)$ relaxes the assumption of a common random effects variance; the variance of the random component associated with subject j in cluster i is $f^2(z_{ij};\gamma)\sigma^2$, so the function f serves to explain some of the extra heterogeneity in the data, over and above that which can be captured in the linear predictor $x'_{ij}\beta$. In the introductory example, for instance, the covariate z_i was not related to the mean response in school i, but instrumental in describing the overdispersion. Model (5.2.1) gives us a straightforward framework for assessing the impact of covariates thought to be associated with the correlation structure in the data. In letting $f(z_{ij};\gamma)$ depend on parameter(s) γ we are able to use the data to estimate the best fitting relationship between z_{ij} and b_i, (for a given family f, and conditioning on x_{ij}).

The parameters (β,γ) from (5.2.1) can be estimated using empirical Bayes methods, or by maximizing the marginal likelihood (3.6.3). The computational burden involved in this latter approach is quite manageable, certainly with a univariate mixing distribution; we shall consider maximum likelihood estimation throughout this chapter.

5.2.2 Related Work

The recent literature reports numerous applications of various random effects models, for binary data in particular. Smyth (1989) describes a generalization of GLMs in which the dispersion parameter may be allowed to depend on covariates in the same manner as does the mean. This leads to two submodels for which an alternating estimation procedure is proposed. Our approach has a similar spirit, but we model the variance of a random effects distribution instead of the dispersion parameter; only in linear models do these coincide. Estimation seems simpler to carry out for (5.2.1) since only a single model is specified. Neuhaus et al. (1992) examine the performance of mixed-effects logistic regression analysis

when the mixture distribution is misspecified. The authors consider situations in which a single mixture distribution applies to all clusters, whether this be the correct or an erroneously assumed distribution. Related to this is the work of Fattinger et al. (1995), in the context of non-linear models. In section 5.5 we investigate the effect of misspecification when the true mixing distribution depends on cluster or individual-level covariates. Our emphasis is not so much on assessing the exact form of the mixing distribution, but rather on establishing a relationship between this distribution and covariates thought to affect intra-cluster correlation. (The connection to the work of Neuhaus et al. (1992) is discussed further in section 5.6). Follman and Wu (1995) present a class of random effects models to deal with missing data in longitudinal studies, an issue which is not addressed here. Cook and Ng (1997) describe an application to disease-state data in which the correlation structure is addressed through a bivariate random effects distribution wherein the two random effects may be correlated. A general probit-normal model which also allows for correlated random effects is proposed by Chan and Kuk (1997); these authors employ an EM algorithm to obtain maximum likelihood estimates, treating the random effects as the missing data component; see also Walker (1996) for a similar application of the EM algorithm.

5.3 Moving Between Cluster-Specific and Population-Averaged Formulations

In this section we focus on the relationship between cluster-specific and population-averaged models, and highlight the notion that each type of model can be thought of as equivalent to a certain formulation of the other type. We do this to emphasize that the marginal correlation structure in a data set can also be modelled by fitting a cluster-specific random effects model. We begin by considering the case of the linear model, and then discuss the

implications in models for binary data.

5.3.1 The Linear Model

Consider the following linear random effects model for the continuous response $\mathcal{Y}_{ij}$ of individual j in cluster i:

$$\mathcal{Y}_{ij} = \mu_{ij} + f_{ij}b_i + \varepsilon_{ij}, \qquad b_i \sim N(0, \sigma^2), \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \tag{5.3.1}$$

where $\mathcal{Y}_{ij}$ and $\mathcal{Y}_{ij'}$ are independent given b_i, and $\{b_i\}$ and $\{\varepsilon_{ij}\}$ are assumed to be i.i.d. samples from their respective distributions. (In the notation of (5.2.1) we would have $\mu_{ij} = x_{ij}'\beta$ and $f_{ij} = f(z_{ij};\gamma)$). Under this model the marginal mean and variance of $\mathcal{Y}_{ij}$ are

$$\mathrm{E}(\mathcal{Y}_{ij}) = \mu_{ij} \qquad \text{and} \qquad \mathrm{Var}(\mathcal{Y}_{ij}) = f_{ij}^2\sigma^2 + \sigma_\varepsilon^2$$

and the marginal correlation is given by

$$\begin{aligned}
\mathrm{Corr}(\mathcal{Y}_{ij}, \mathcal{Y}_{ij'}) &= \frac{\mathrm{Cov}(\mathcal{Y}_{ij}, \mathcal{Y}_{ij'})}{\sqrt{\mathrm{Var}(\mathcal{Y}_{ij})\mathrm{Var}(\mathcal{Y}_{ij'})}} \\
&= \frac{\mathrm{E}_b\{\mathrm{Cov}(\mathcal{Y}_{ij}, \mathcal{Y}_{ij'}|\ b_i)\} + \mathrm{Cov}_b\{\mathrm{E}(\mathcal{Y}_{ij}|\ b_i), \mathrm{E}(\mathcal{Y}_{ij'}|\ b_i)\}}{\sqrt{(f_{ij}^2\sigma^2 + \sigma_\varepsilon^2)(f_{ij'}^2\sigma^2 + \sigma_\varepsilon^2)}} \\
&= \frac{f_{ij}f_{ij'}\sigma^2}{\sqrt{(f_{ij}^2\sigma^2 + \sigma_\varepsilon^2)(f_{ij'}^2\sigma^2 + \sigma_\varepsilon^2)}} \quad \left(= \frac{f_i^2\sigma^2}{f_i^2\sigma^2 + \sigma_\varepsilon^2} \text{ if } f_{ij} = f_i\ \forall\ j\right).
\end{aligned}$$

Model 5.3.1 is therefore equivalent to the marginal model

$$\mathcal{Y}_{ij} = \mu_{ij} + v_{ij}, \qquad \mathrm{Corr}(\mathcal{Y}_{ij}, \mathcal{Y}_{ij'}) = \frac{f_{ij}f_{ij'}\sigma^2}{\sigma_{ij}\sigma_{ij'}}, \tag{5.3.2}$$

where $v_{ij} \sim N(0, \sigma_{ij}^2)$, $\mathrm{Cov}(v_{ij}, v_{ij'}) = f_{ij}f_{ij'}\sigma^2$ and $\sigma_{ij}^2 = f_{ij}^2\sigma^2 + \sigma_\varepsilon^2$; (observations from different clusters are independent, as before). Note that as the random effects variance

σ^2 approaches 0, $\text{Corr}(\mathcal{Y}_{ij}, \mathcal{Y}_{ij'}) \to 0$ and as $\sigma^2 \to \infty$, $\text{Corr}(\mathcal{Y}_{ij}, \mathcal{Y}_{ij'}) \to 1$ as one should expect. Thus the random effects model (5.3.1) represents one way to accomodate intracluster correlation and happens in this case to correspond to a marginal model with unequal variances.

Alternatively, a population-averaged model assuming equal marginal variances and an arbitrary intracluster correlation structure cannot be expressed, in general, in the form (5.3.1). Consider for instance the model

$$\mathcal{Y}_{ij} = \mu_{ij} + \epsilon_{ij}, \qquad \text{Corr}(\mathcal{Y}_{ij}, \mathcal{Y}_{ij'}) = \rho_{ijj'}, \tag{5.3.3}$$

where $\text{Var}(\epsilon_{ij}) = \text{Var}(\mathcal{Y}_{ij}) = \nu^2$. To write (5.3.3) in the form of (5.3.1) one would need to fix f_{ij} at a constant value. This in turn would also imply a fixed value for $\rho_{ijj'}$.

Summarizing, in models (5.3.1) and (5.3.2) the marginal variances are functions of f_{ij} and $f_{ij'}$, as are the marginal correlations. In linear models in general one can also specify constant marginal variances, and arbitrary intracluster correlations - the case in (5.3.3). This distinction fades however when considering generalized linear models for binary data, in which the variance is already a function of the mean and therefore specific to each observation. We shall discuss this case next.

5.3.2 Marginal Correlation Estimates Induced by Random Effects Models for Binary Data

Model (5.2.1) is a cluster-specific formulation for the conditional probability

$$p_{ij} = P(Y_{ij} = 1 |\ x_{ij}, z_{ij}, b_i).$$

As Neuhaus et al. (1991) point out, this also specifies a unique marginal model for Y_{ij} (averaging over the distribution $g(b_i)$ of b_i - $N(0, \sigma^2)$ in this case):

$$P(Y_{ij} = 1|\ x_{ij}, z_{ij}) = \int P(Y_{ij} = 1|\ x_{ij}, z_{ij}, b_i) g(b_i) db_i. \tag{5.3.4}$$

(See also the discussion in section 3.2). In the same manner we can obtain the marginal correlation between Y_{ij} and $Y_{ij'}$ induced by (5.2.1). Noting that

$$\begin{aligned}
\mathrm{Cov}(Y_{ij}, Y_{ij'}|\ x, z) &= \mathrm{E}_b\{\mathrm{Cov}(Y_{ij}, Y_{ij'}|\ x, z, b_i)\} + \\
&\qquad \mathrm{Cov}_b\{\mathrm{E}(Y_{ij}|\ x, z, b_i), \mathrm{E}(Y_{ij'}|\ x, z, b_i)\} \\
&= 0 + \mathrm{Cov}_b(p_{ij}, p_{ij'}) \\
\text{and} \quad \mathrm{Var}(Y_{ij}|\ x, z) &= \mathrm{E}(Y_{ij}^2|\ x, z) - [\mathrm{E}(Y_{ij}|\ x, z)]^2 \\
&= \mathrm{E}_b\{\mathrm{E}(Y_{ij}^2|\ x, z, b_i)\} - [\mathrm{E}_b(p_{ij})]^2 \\
&= \mathrm{E}_b\{p_{ij}(1 - p_{ij}) + p_{ij}^2\} - [\mathrm{E}_b(p_{ij})]^2 \\
&= \mathrm{E}_b(p_{ij})(1 - \mathrm{E}_b(p_{ij})),
\end{aligned}$$

we have

$$\begin{aligned}
\mathrm{Corr}(Y_{ij}, Y_{ij'}|\ x, z) &= \frac{\mathrm{Cov}(Y_{ij}, Y_{ij'}|\ x, z)}{\sqrt{\mathrm{Var}(Y_{ij}|\ x, z)\mathrm{Var}(Y_{ij'}|\ x, z)}} \\
&= \frac{\mathrm{Cov}_b(p_{ij}, p_{ij'})}{\sqrt{\mathrm{E}_b(p_{ij})(1 - \mathrm{E}_b(p_{ij}))\mathrm{E}_b(p_{ij'})(1 - \mathrm{E}_b(p_{ij'}))}}
\end{aligned} \tag{5.3.5}$$

In the conditional expressions above, x and z are used as generic variables to indicate conditioning on whichever covariates are appropriate. Thus for example $\mathrm{Corr}(Y_{ij}, Y_{ij'}|\ x, z) = \mathrm{Corr}(Y_{ij}, Y_{ij'}|\ x_{ij}, x_{ij'}, z_{ij}, z_{ij'})$. Closed-form expressions exist neither for (5.3.4) nor for (5.3.5). However the approximation due to Johnson and Kotz (1970) (given in equation

(3.5.23)) is frequently used to obtain the fairly accurate estimate

$$P(Y_{ij}=1|\ x,z) \quad = \quad \mathrm{E}_b(p_{ij}) \quad \simeq \quad \frac{e^{x'_{ij}\beta/\sqrt{1+c^2 f^2(z_{ij};\gamma)\sigma^2}}}{1+e^{x'_{ij}\beta/\sqrt{1+c^2 f^2(z_{ij};\gamma)\sigma^2}}}, \tag{5.3.6}$$

$c = 16\sqrt{3}/15\pi$. Similarly one can also derive an approximation for the marginal correlation given by (5.3.5). We develop this for two cases, the first assuming covariates on the level of the cluster only, the second allowing for individual-level covariates.

For the first case we consider the model

$$\begin{aligned} Y_{ij}|\ b_i &\sim \mathrm{Bin}(1,p_i), \qquad b_i \sim N(0,\sigma^2) \\ p_i &= p_i(b_i) = \frac{e^{x'_i\beta+f(z_i;\gamma)\cdot b_i}}{1+e^{x'_i\beta+f(z_i;\gamma)\cdot b_i}}. \end{aligned} \tag{5.3.7}$$

In this case

$$\mathrm{Corr}(Y_{ij},Y_{ij'}|\ x,z) = \frac{\mathrm{Var}_b(p_i)}{\mathrm{E}_b(p_i(1-p_i))+\mathrm{Var}_b(p_i)}. \tag{5.3.8}$$

Using a second-order Taylor series expansion of $p_i(b_i)$ around $b_i = 0$ leads to the approximation

$$\mathrm{Var}_b(p_i) \simeq \pi_i^2(1-\pi_i)^2(1+\frac{(1-2\pi_i)^2 f_i^2\sigma^2}{2})f_i^2\sigma^2,$$

and a similar expansion of $p_i(b_i)(1-p_i(b_i))$ gives

$$\mathrm{E}_b(p_i(1-p_i))+\mathrm{Var}_b(p_i) \simeq \pi_i(1-\pi_i)[1+\frac{f_i^2\sigma^2}{2}-2\pi_i(1-\pi_i)f_i^2\sigma^2+\frac{\pi_i(1-\pi_i)(1-2\pi_i)^2}{2}f_i^4\sigma^4],$$

where $\pi_i = p_i(0)$ and $f_i = f(z_i;\gamma)$. Therefore (5.3.8) can be approximated by

$$\rho_i \quad = \quad \mathrm{Corr}(Y_{ij},Y_{ij'}|\ x,z)$$

$$\simeq \frac{\pi_i(1-\pi_i)[1+\frac{(1-2\pi_i)^2 f_i^2\sigma^2}{2}]f_i^2\sigma^2}{1+\frac{f_i^2\sigma^2}{2}-2\pi_i(1-\pi_i)f_i^2\sigma^2+\frac{\pi_i(1-\pi_i)(1-2\pi_i)^2}{2}f_i^4\sigma^4}. \tag{5.3.9}$$

Note that as in the linear case, for fixed π_i, as $f_i^2\sigma^2 \longrightarrow \infty$, the approximation for ρ_i approaches 1, and as $f_i^2\sigma^2 \longrightarrow 0$ it approaches 0, as expected. Equation (5.3.9) emphasizes the flexibility which the function $f(z_i;\gamma)$ provides in model (5.3.7): for fixed π_i and σ^2, varying degrees of intra-cluster correlation can be accomodated through the cluster dependent value of f_i. Table 5.2 compares the performance of the approximation for ρ_i with the exact value, obtained from (5.3.8). Specifically, letting $\varphi(b,\sigma^2)$ denote the $N(0,\sigma^2)$ density, we used numerical integration to evaluate the integrals

$$\mathrm{E}_b(p_i(b_i)) = \int p_i(b)\varphi(b,\sigma^2)db \quad \text{and} \quad \mathrm{E}_b(p_i^2(b_i)) = \int p_i^2(b)\varphi(b,\sigma^2)db$$

and using these, evaluated equation (5.3.8). The top row in each pair of rows in table 5.2 gives the approximate value, to be compared to the exact value given in the bottom row. For combinations of σ^2 and f_i in which the product $f_i^2\sigma^2$ is not too large, say less than or equal to 1.5, approximation (5.3.9) is reasonably close to the true value of ρ_i. (We note in passing that this restriction on f_i and σ^2 is moderate enough to cover many situations of practical interest). Observe nevertheless that the true value of the marginal correlation tends to be less sensitive to the value of π_i than the approximation to it suggests; this tends to underestimate ρ_i for π_i near 0 or 1, but overestimate it for π_i near 0.5. In addition, whereas (5.3.9) will never produce a negative estimate, a desirable feature when analyzing overdispersed data, it can for some combinations of f_i and σ^2 take on values larger than 1, as seen in table 5.2. In such regions the approximation is of course not helpful.

$\sigma^2 = 0.1$					π_i			
f_i		0.05	0.10	0.25	0.50	0.75	0.90	0.95
1	(A)	0.00475	0.00900	0.0187	0.0250	0.0187	0.00900	0.00475
	(E)	0.00508	0.00938	0.0185	0.0238	0.0185	0.00938	0.00508
2	(A)	0.0189	0.0359	0.0747	0.1000	0.0747	0.0359	0.0189
	(E)	0.0242	0.0410	0.0700	0.0842	0.0700	0.0410	0.0242
3	(A)	0.0423	0.0796	0.1659	0.2250	0.1659	0.0796	0.0423
	(E)	0.0658	0.0979	0.1421	0.1607	0.1421	0.0979	0.0658
4	(A)	0.0738	0.1373	0.2857	0.4000	0.2857	0.1373	0.0738
	(E)	0.1312	0.1722	0.2203	0.2386	0.2203	0.1722	0.1312
$\sigma^2 = 0.2$					π_i			
f_i		0.05	0.10	0.25	0.50	0.75	0.90	0.95
1	(A)	0.00949	0.0180	0.0375	0.0500	0.0375	0.0180	0.00949
	(E)	0.0108	0.0194	0.0363	0.0456	0.0363	0.0194	0.0108
2	(A)	0.0376	0.0710	0.1480	0.2000	0.1480	0.0710	0.0376
	(E)	0.0568	0.0865	0.1289	0.1471	0.1289	0.0865	0.0568
3	(A)	0.0825	0.1529	0.3178	0.4500	0.3178	0.1529	0.0825
	(E)	0.1495	0.1913	0.2390	0.2568	0.2390	0.1913	0.1495
4	(A)	0.1400	0.2514	0.5122	0.8000	0.5122	0.2514	0.1400
	(E)	0.2615	0.3002	0.3401	0.3540	0.3401	0.3002	0.2615
$\sigma^2 = 0.5$					π_i			
f_i		0.05	0.10	0.25	0.50	0.75	0.90	0.95
1	(A)	0.0237	0.0447	0.0932	0.1250	0.0932	0.0447	0.0237
	(E)	0.0317	0.0522	0.0857	0.1014	0.0857	0.0522	0.0317
2	(A)	0.0911	0.1682	0.3488	0.5000	0.3488	0.1682	0.0911
	(E)	0.1673	0.2094	0.2563	0.2736	0.2563	0.2094	0.1673
3	(A)	0.1878	0.3269	0.6472	1.1250	0.6472	0.3269	0.1878
	(E)	0.3395	0.3722	0.4046	0.4157	0.4046	0.3722	0.3395
4	(A)	0.2945	0.4744	0.8571	2.0000	0.8571	0.4744	0.2945
	(E)	0.4728	0.4935	0.5133	0.5199	0.5133	0.4935	0.4728

(continued)

$\sigma^2 = 1.0$		π_i						
f_i		0.05	0.10	0.25	0.50	0.75	0.90	0.95
1	(A)	0.0469	0.0881	0.1837	0.2500	0.1837	0.0881	0.0469
	(E)	0.0750	0.1091	0.1547	0.1735	0.1547	0.1091	0.0750
2	(A)	0.1700	0.2995	0.6000	1.0000	0.6000	0.2995	0.1700
	(E)	0.3121	0.3471	0.3821	0.3942	0.3821	0.3471	0.3121
3	(A)	0.3201	0.5059	0.8913	2.2500	0.8913	0.5059	0.3201
	(E)	0.4987	0.5171	0.5347	0.5406	0.5347	0.5171	0.4987
4	(A)	0.4583	0.6531	1.0000	4.0000	1.0000	0.6531	0.4583
	(E)	0.6133	0.6230	0.6322	0.6352	0.6322	0.6230	0.6133

Table 5.2: Performance of the approximation to $\rho_i = \mathrm{Corr}(Y_{ij}, Y_{ij'} |\, x, z)$; rows flagged by (A) and (E) contain the approximate and exact values of ρ_i respectively, for the given combinations of π_i, f_i and σ^2.

In a similar fashion, the marginal correlation (5.3.5) can be approximated when covariates are also on the level of the individual. In this case the model of interest is

$$\begin{aligned} Y_{ij} |\, b_i &\sim \mathrm{Bin}(1, p_{ij}), \qquad b_i \sim N(0, \sigma^2) \\ \log \frac{p_{ij}}{1 - p_{ij}} &= x_{ij}'\beta + f(z_{ij}; \gamma) \cdot b_i. \end{aligned} \tag{5.3.10}$$

Writing p_{ij} as $p_{ij}(b_i)$ and using Taylor series expansions as above leads to the following approximation:

$$\begin{aligned} \rho_{ijj'} &= \mathrm{Corr}(Y_{ij}, Y_{ij'} |\, x, z) \\ &= \sqrt{\frac{\pi_{ij}(1-\pi_{ij})\pi_{ij'}(1-\pi_{ij'})}{v_{ij}v_{ij'}}} \times \\ &\qquad \{1 + \frac{(1-2\pi_{ij})(1-2\pi_{ij'})f_{ij}f_{ij'}\sigma^2}{2}\} f_{ij}f_{ij'}\sigma^2, \end{aligned} \tag{5.3.11}$$

where

$$v_{ij} = 1 + \frac{f_{ij}^2\sigma^2}{2} - 2\pi_{ij}(1-\pi_{ij})f_{ij}^2\sigma^2 + \frac{1}{2}\pi_{ij}(1-\pi_{ij})[1-2\pi_{ij}]^2 f_{ij}^4\sigma^4$$

and $\pi_{ij} = p_{ij}(0)$, $f_{ij} = f(z_{ij};\gamma)$.

5.3.3 Random Effects Variance Estimates Induced by Marginal Models for Binary Data

So far the discussion has focussed on the transition from a cluster-specific model, of the form (5.3.7) or (5.3.10), to the population-averaged model induced by it. Conversely, one might begin with a marginal model formulation and determine a cluster-specific equivalent. As noted by Neuhaus et al. (1991), in the first case a unique marginal is implied, but a given marginal model does not specify any single mixed-effects model. Let us therefore restrict our attention to the class of logistic-normal random (mixed) effects models in determining a cluster-specific model arising as an equivalent analogue to a marginal model.

We can describe a marginal, or population-averaged model for binary data by its marginal probabilities p_M and pairwise correlations ρ. Each of these may be thought of as functions of cluster-specific probabilities π (with π_i depending on f_i) and the variance σ^2 of a random effects distribution, used to specify a logistic-normal mixed effects model. The relationship between (p_M, ρ) and (π, σ^2) can be explicitly investigated, for example using the approximations given above. In the case of cluster-level covariates only, for instance, setting $p_{Mi} = P(Y_{ij} = 1|\ x, z)$ and $\rho_i = \mathrm{Corr}(Y_{ij}, Y_{ij'}|\ x, z)$ equal to the expressions in (5.3.6) and (5.3.9) respectively, yields two equations in two unknowns which may be solved for π_i and σ^2; (recall that $\pi_i = p_i(0)$, referring to (5.3.7)). One obtains the solutions

$$\pi_i = \frac{e^{\log(p_{Mi}/(1-p_{Mi}))\sqrt{1+c^2 f_i^2 \sigma^2}}}{1 + e^{\log(p_{Mi}/(1-p_{Mi}))\sqrt{1+c^2 f_i^2 \sigma^2}}} \tag{5.3.12}$$

and

$$\sigma^2 = \begin{cases} 4\rho_i/f_i^2 & \text{if } \pi_i = 0.5 \\ \frac{-d_i - \sqrt{d_i^2 - 4a_i c_i}}{2f_i^2 a_i}, & \text{if } \pi_i \neq 0.5, \end{cases} \tag{5.3.13}$$

where

$$a_i = -\frac{1}{2}(1-\rho_i)\pi_i(1-\pi_i)(1-2\pi_i)^2,$$

$$d_i = (\frac{1}{2} - 2\pi_i(1-\pi_i))\rho_i - \pi_i(1-\pi_i), \qquad c_i = \rho_i$$

It is of interest in particular to study the behaviour of σ^2 as a function of p_{Mi} and ρ_i. Figure 5.1 depicts this graphically, plotting the estimate of σ^2 given in (5.3.13) against a grid of values corresponding to combinations of p_{Mi} and ρ_i. Note that we have assumed

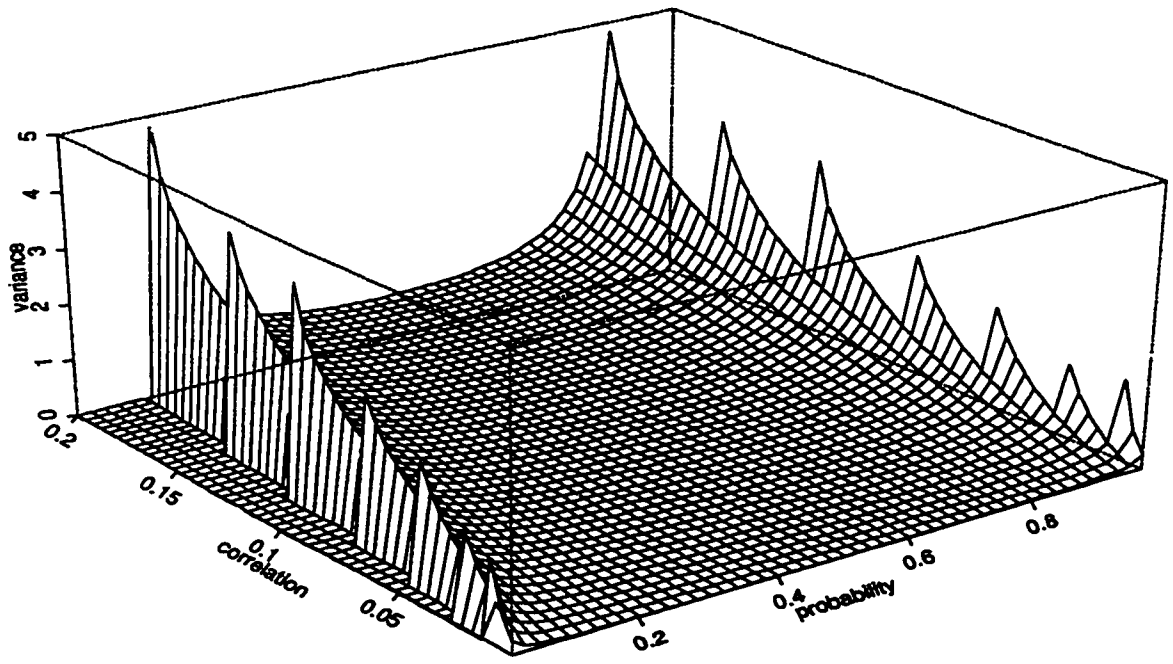


Figure 5.1: Perspective plot of σ^2 vs (p_{Mi}, ρ_i).

here that $f_i = 1$; alternatively one can interpret this graph as plotting $f_i^2\sigma^2$ vs (p_{Mi}, ρ_i). Clearly σ^2 is an increasing function of intra-cluster correlation, and for any fixed value of ρ_i, the more extreme the marginal probability, the larger the random effects variance implied. This is the sort of behaviour one would expect; under a cluster probability close to 0 or 1, most observations tend already to be of the same type, successes or failures. Introducing positive correlation among the observations in addition serves to emphasize this tendency further, making it necessary to postulate perhaps a very large random effect to account for the extreme response rate.

5.3.4 Example: the Beta-Binomial Model

Here we draw a brief comparison between the beta-binomial model and the random effects model (5.3.7). The former is flexible enough to allow a cluster-dependent modelling of the correlation structure in the data, independent of the mean, whereas the latter will admit a cluster-dependent estimate of the random effects variance. The above discussion suggests that we should therefore be able to find such a random effects model that closely resembles the beta-binomial. We can restate the problem by focussing on a single cluster, fixing the two parameters of the beta-binomial distribution (giving us a marginal probability and correlation pair (p_M, ρ)), computing an appropriate random effects variance estimate and a probability for the cluster-specific model, and comparing the resulting probability distribution to the beta-binomial. To fix ideas, suppose that $R = \sum_{j=1}^{n} Y_j$ is distributed as beta-binomial(a, b), so that $p_M = \mathrm{E}(Y_j) = a/(a+b)$, $\rho = \mathrm{Corr}(Y_j, Y_{j'}) = (1+a+b)^{-1}$ and

$$P(R = r) = \frac{(a+r-1)^{(r)}(b+n-r-1)^{(n-r)}}{(a+b+n-1)^{(n)}}, \qquad r = 1, 2, ..., n. \tag{5.3.14}$$

(See also equation (3.4.3)). We would like to compare this probability function with the analogous one produced by the random effects model, that being

$$P(R=r)=\int_{-\infty}^{\infty}\binom{n}{r}p(b)^r(1-p(b))^{n-r}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-b^2/2\sigma^2}db,\quad r=1,...,n. \qquad (5.3.15)$$

We can use equations (5.3.12) and (5.3.13) to approximate $p(b)$ and σ^2, respectively; in this case

$$p(b)=\frac{e^{\log(a/b)\sqrt{1+c^2f^2\sigma^2}+f\cdot b}}{1+e^{\log(a/b)\sqrt{1+c^2f^2\sigma^2}+f\cdot b}}$$

with $\pi_i = \pi = p(0)$, from which the approximation for σ^2 easily follows. Figure 5.2 compares the probability histograms for the beta-binomial distribution (5.3.14) to those of the approximating distribution (5.3.15), based on the random effects model, for $(a,b) = (0.9,8.1)$, (1.8,7.2) and (4.5,4.5) (i.e. $p_M = 0.1$, 0.2 and 0.5 and $\rho = 0.1$). Assume without loss of generality that $f = 1$. The binomial probability histograms, assuming $\rho = 0$, are also shown for comparison. In all cases we have considered a cluster size of $n = 50$, for which an intra-cluster correlation of $\rho = 0.1$ is very substantial, leading to a variance inflation factor of almost 6. The distribution (5.3.15) approximates (5.3.14) very well for all 3 marginal probabilities, capturing the effect of overdispersion in the same manner as the beta-binomial model does. The close correspondence between the two models is further emphasized by comparing their histograms to the ones based on the binomial model, assuming nominal dispersion.

We have therefore found a distribution based on a cluster-specific random effects model which is very similar to that based on the marginal beta-binomial model. Just as in modelling the responses from several clusters using the beta-binomial model we can allow intra-cluster correlation to be cluster-dependent, so too we can achieve a similar fit by introducing an appropriate cluster-specific function $f(z_i;\gamma)$ into a random effects model, as in (5.3.7). Further discussion on this is provided in the next section. There are several

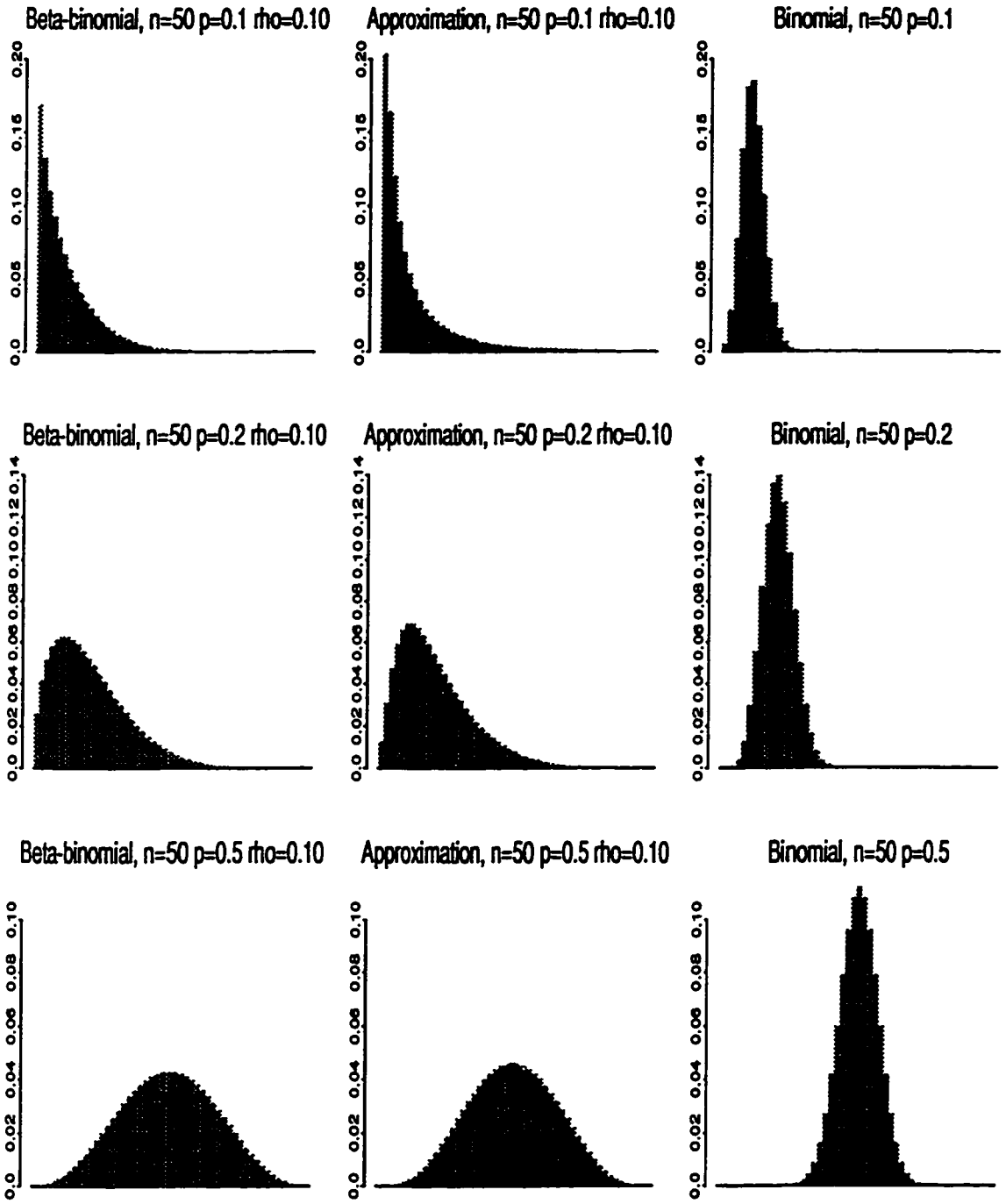


Figure 5.2: Graphical examination of a random effects model approximating the beta-binomial; the bars in each plot represent the probabilities $P(R = r)$, for r ranging from 0 to 50. The binomial probability histograms are shown for comparison.

advantages of this latter modelling approach over the former: it provides a more natural framework for regression analysis; it allows one to retain the context of the logistic model; and perhaps most importantly, it extends easily to situations involving individual-level covariates, which the beta-binomial model cannot accomodate.

5.4 The Function $f(z;\gamma)$

5.4.1 Cluster-Level Covariates

In this section we consider cases where the function $f(z;\gamma)$ depends on a cluster-level covariate z. A situation in which z varies from individual to individual within a cluster is discussed in the next section.

It would be desirable at the stage of exploratory analysis to have a simple means of assessing whether or not much can be gained through modelling the random cluster effect as $f(z_i;\gamma)\cdot b_i$, $b_i \sim N(0,\sigma^2)$, and given this, what functional forms for $f(z_i;\gamma)$ would be most appropriate. The first of these objectives can be addressed to a limited extent by simple plotting techniques, whereas the second is very difficult but also less important to meet.

Consider model (5.3.7), with r_i/n_i the observed proportion of successes in the ith cluster, $i = 1,...,K$. The relationship between the random cluster effect and the covariate of interest z should be roughly indicated by a plot of

$$\hat{\varepsilon}_i = \log(\frac{r_i}{n_i - r_i}) - x_i'\hat{\beta} \quad \text{vs} \quad z_i, \qquad i = 1,...,K, \tag{5.4.1}$$

where $\hat{\beta}$ is the maximum likelihood estimate of β from a standard logistic model fit to the

data. The residuals $\hat{\epsilon}_i$ represent the discrepancies on a logistic scale between the observed and the fitted proportions, i.e. the errors which the random effects are trying to explain. The plot $\hat{\epsilon}_i$ vs z_i may therefore be interpreted in a similar manner as a residual plot for a linear model. The usual precautions can be applied in (5.4.1) to avoid infinite values, that is, replacing $r_i/(n_i - r_i)$ by $(r_i + 0.5)/(n_i - r_i + 0.5)$ whenever $r_i = 0$ or n_i. Furthermore if the fixed effects covariates include ones on the level of the individual, one might replace $x_i'\hat{\beta}$ with $\sum_j x_{ij}'\hat{\beta}/n_i$.

Unfortunately experience shows that the random effects variance must depend rather heavily on the covariate z for a plot of ϵ_i vs z_i to show a clear pattern. Of course this also depends on the number of clusters available, and to some extent on the cluster size. If the covariate of interest is cluster size, then the range of cluster sizes will play a crucial role. It may be useful to plot $\hat{\epsilon}_i^2$ vs z_i in addition to (5.4.1), since $\hat{\epsilon}_i^2$ is more directly related to the random effects variance. Another alternative is to fit a standard empirical Bayes random effects model (assuming a constant random effects variance) and plot $\widehat{\text{Var}}(\hat{b}_i)$ vs z_i, where $\hat{b}_i$ is the random effects estimate for cluster i. One should expect a violation of the assumption of constant variance to be reflected in the estimates of variability associated with the random effects estimates. However one must be careful here especially when the covariate of interest z is cluster size, since $\widehat{\text{Var}}(\hat{b}_i)$ is already inherently a function of cluster size.

Exactly how large an effect is required in order to be able to detect it at a given Type I error rate in the model fitting process is discussed in section 5.5. Here we give a few examples from our experience of visual examination of the plots discussed above. Consider the model $\text{logit}(p_i) = u_i^7 b_i$, where $b_i \sim N(0,1)$ and $u_i \sim \text{Unif}(0,1)$. Since we are only interested in the random component of the linear predictor here, we omit specifying a model for the mean response beyond a common mean of 0.5 for all clusters. For $\gamma = 0.25$, for instance, as many as 200 clusters would be required to detect an increase in the magnitude of ϵ_i with u_i. For $\gamma \geq 0.5$, 50 clusters seem to be sufficient. Consider a similar model,

$\text{logit}(p_i) = e^{\gamma u_i} b_i$; detecting an increasing pattern in the residuals at all suggests that γ is more than likely greater than 0.5, and even for $\gamma = 0.5$ at least 100 clusters seem to be required. The more convex the function $f(z_i;\gamma) = f(u_i;\gamma)$ is, the more difficult it is to interpret the residual plot, especially if K is small. A wider range of u_i or a larger value of the variance of b_i would of course allow us to detect a pattern in the residuals for smaller values of γ, and with fewer clusters; see also section 5.5. We considered as well the case where the magnitude of the random effects is a decreasing function of cluster size. Specifically, consider the model $\text{logit}(p_i) = b_i/(n_i)^\gamma$. For $\gamma = 0.25$ and n_i ranging linearly from 20 to 200, a decreasing trend in the residuals is evident with about 50 clusters or more. For $\gamma \geq 0.5$ one should be able to detect a pattern with a range of cluster sizes as small as 20 to 100. This is assuming in each case that $b_i \sim N(0, \sigma^2)$, where σ^2 is chosen so that the random effects variance in the smallest cluster equals 2; i.e. $\sigma^2/20^{2\gamma} = 2$.

In all cases it seems that one should not look to such residual plots for direction unless one has data on a fair number of clusters, i.e. 50 at the very least. They are useful at best for identifying general relationships, such as increasing or decreasing variablility with a covariate of interest. In most cases we are only looking to detect this kind of simple relationship, and the precise form of $f(z_i;\gamma)$ is almost immaterial. Consider for example two models for the random effect $f(z_i;\gamma)b_i$:

$$\begin{array}{lll} i) & f_{1i}b_i = z_i^{\gamma_1} b_i, & \text{vs} \qquad ii) \quad f_{2i}b_i = e^{\gamma_2 z_i} b_i, \\ & b_i \sim N(0, \sigma_1^2) & \qquad\qquad\quad b_i \sim N(0, \sigma_2^2), \end{array}$$

where $z_i \sim \text{Unif}(a, b)$, $b > a > 0$. We can easily determine values γ_1, σ_1^2, γ_2 and σ_2^2 which will induce similar behaviour in $f_{1i}b_i$ and $f_{2i}b_i$, suggesting that misspecification of one function by choosing the other will not have have a dramatic effect. For instance, note that the variance of $f_{1i}b_i$ ranges in $(a^{2\gamma_1}\sigma_1^2, b^{2\gamma_1}\sigma_1^2) = (e^{l_1}, e^{u_1})$, and that of $f_{2i}b_i$ in $(e^{2\gamma_2 a}\sigma_2^2, e^{2\gamma_2 b}\sigma_2^2) = (e^{l_2}, e^{u_2})$. The two functions cover the same range in their variances,

with $l_1 = l_2 = l$ and $u_1 = u_2 = u$, if

$$\gamma_1 = \frac{l-u}{2(\log(a)-\log(b))}, \quad \sigma_1^2 = \exp\{u - \frac{(l-u)\log(b)}{\log(a)-\log(b)}\},$$
$$\gamma_2 = \frac{l-u}{2(a-b)}, \quad \sigma_1^2 = \exp\{l - \frac{l-u}{a-b}a\}.$$

The only difference between them is that $f_{2i}b_i$ is more convex than $f_{1i}b_i$, which in most cases is unlikely to make a substantial difference to the fit of a model. A vast amount of data would be needed to distinguish between $f_{1i}b_i$ and $f_{2i}b_i$; in finite samples both functional forms will capture the same qualitative trend.

5.4.2 Individual-Level Covariates

In some cases intracluster correlation may be a function of individual-level covariates, and this can also be accomodated through a random effects model, of the form (5.3.10). We shall consider specifically cases where the correlation is a function of a categorical variable, with pairs of individuals having the same value of this variable sharing a similar correlation depending on this value. This has direct relevance to the WSPP3: the data collected on each student includes a risk score, categorized as high, medium or low, giving an indication of the student's *individual* risk of smoking. It is of interest to determine whether or not there is additional variability in the data, after accounting for random school effects, which is due to students' individual risk profiles. This information could help explain the school-to-school variability in smoking rates and provide some direction as to where one should concentrate the efforts of smoking prevention programs.

We suggest a simple graphical tool which is helpful in detecting such varying dispersion among subgroups within clusters. For the purpose of illustration we shall consider a situation in which each individual belongs to one of two groups within a given cluster, and whose response is a function of a random effect at the level of the cluster multiplied by a

factor depending on the group the individual is in; this allows for varying dispersion in the two groups which cannot be accounted for by the cluster-level random effect alone. (The generalization to more than two groups is straightforward; see Chapter 7). In mathematical terms we can describe the model as

$$\log(\frac{p_{ij}}{1-p_{ij}}) = x_{ij}'\beta + e^{\gamma z_{ij}} b_i, \qquad b_i \sim N(0,\sigma^2), \tag{5.4.2}$$

where $z_{ij} = 0$ or 1, according to whether the jth individual in cluster i belongs to group 1 or 2 respectively. If $\gamma > 0$ there is greater overdispersion (stronger correlation) among individuals in group 2 across clusters than among individuals in group 1, and vice versa if $\gamma < 0$. To detect departures of γ from 0 one might proceed as follows:

1. Fit a standard logistic model to the data to obtain $\hat{\beta}$.

2. For each group, compute the residuals

$$\hat{\varepsilon}_{ig} = \log(\frac{r_{ig}}{n_{ig}-r_{ig}}) - \sum_{j|z_{ij}=g} x_{ij}'\hat{\beta}/n_{ig}, \qquad i = 1,...,K, \quad g = 1,2,$$

   where r_{ig}/n_{ig} is the observed proportion of successes in group g.

3. Sort the clusters by overall proportion of successes (r_i/n_i), from smallest to largest, and save the resulting set of cluster indices (the order statistics) in the vector m.

4. On the same graph plot $\hat{\varepsilon}_{ig}$ vs m_i for $g = 1,2$.

If $\gamma > 0$ the residuals $\hat{\varepsilon}_{i2}$ are generally larger in magnitude than $\hat{\varepsilon}_{i1}$, particlarly for clusters with very small or large response proportions. The converse holds if $\gamma < 0$. Figure 5.3 shows some typical patterns one might observe, for data sets with varying numbers of clusters, observations within groups and effect sizes. We generated the data from model (5.4.2) letting $x_{ij}'\beta = 0$ for simplicity, since we are not concerned with the mean specification of

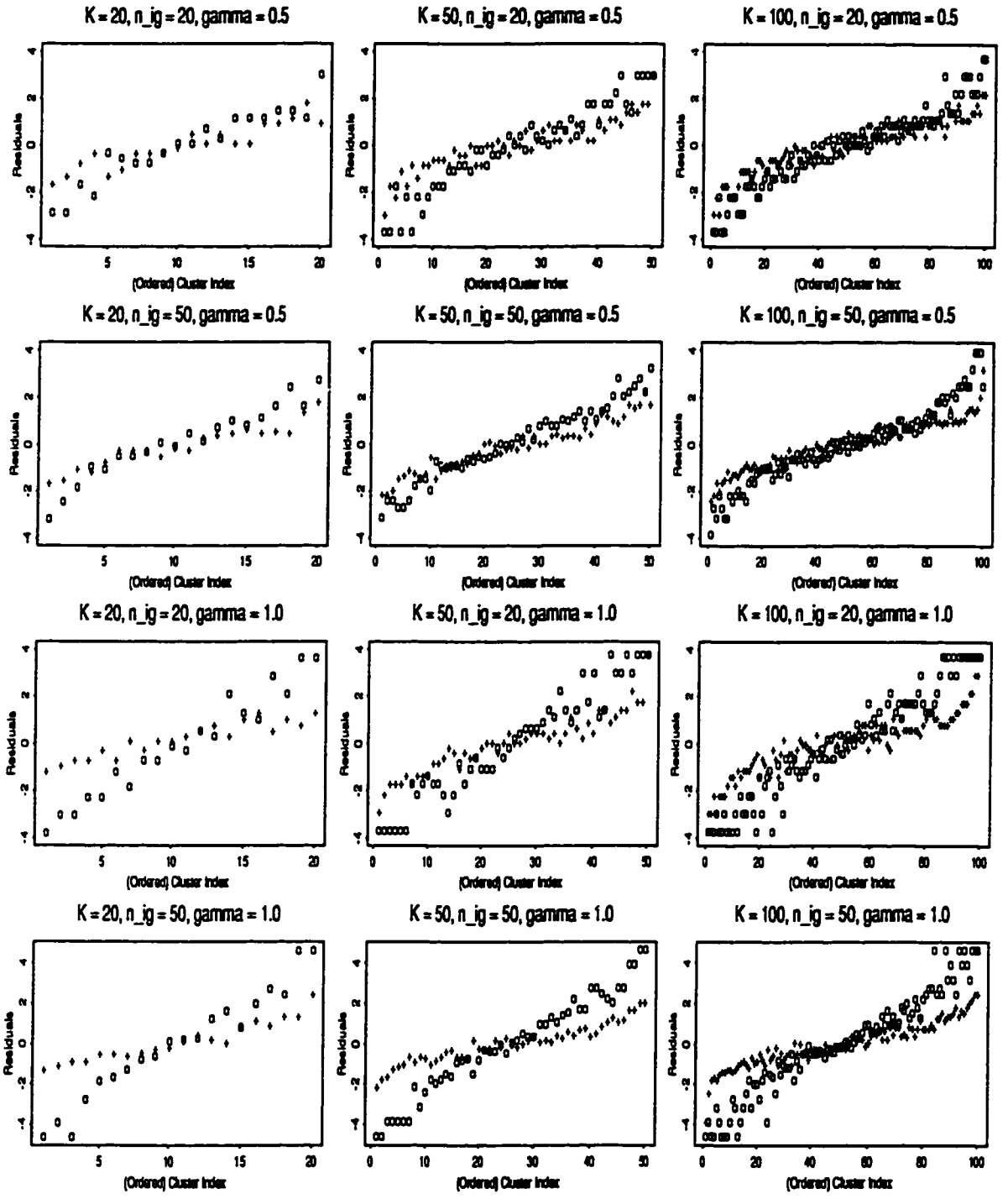


Figure 5.3: Some typical residual plots of $\hat{\epsilon}_{ig}$ vs m_i, $g = 1, 2$, for various values of K, n_{ig} and γ. Observations from group 1 are indicated by a '+', and those from group 2 by 'o'.

the model here. In all cases we chose $\sigma^2 = 1$ and a constant cluster size, with half of the observations within each cluster belonging to group 1, and the other half to group 2.

Examination of such plots is useful even with a smaller number of clusters, our experience suggesting that 20 seem to be sufficient. One drawback is that on average 10 or more observations per group within clusters seem to be required. In each of the plots shown in figure 5.3 it is clear that the residuals belonging to observations from group 2 (marked by 'o's) are larger in magnitude than those from group 1 (marked by '+'s). The residual plot for the combined observations from each cluster lies roughly between the two group-specific plots.

The approach advanced here is a simple but effective way of visually assessing the inadequacy of assuming a single cluster-level random effect, when the departure from this assumption is of the manner described. Note that we are assuming the presence of two random effects distributions in model (5.4.2), both with mean zero but one with variance σ^2, the other with variance $e^{2\gamma}\sigma^2$. Alternatively one could specify a bivariate random effects distribution in this case, but the representation given here is more parsimonious and less computationally intensive in terms of implementation. We have only considered a simple example here, but this latter point becomes increasingly important as the number of levels of z_{ij} increases.

5.5 Some Simulation Results

In this section we address two issues of interest regarding the models discussed. First we investigate how large an effect (in terms of the magnitude of γ in $f(z_{ij};\gamma)$) is required in order to be able to detect it with a certain power at a given type I error rate. Secondly, we shall examine the effect of model misspecification (assuming that $f(z_{ij};\gamma) = 1$, given that the data are generated from a model with $f(z_{ij};\gamma) \neq 1$). These results are not extensive but give some indication of what one might expect from this general class of random effects

models.

5.5.1 The Power for Testing $H_0 : \gamma = 0$ vs $H_A : \gamma > 0$

Assuming that $\gamma = 0$ implies $f(z_{ij};\gamma) = 1$, section 5.4 was concerned with visual assessment of whether or not $\gamma = 0$. In this section we investigate the same question more formally. In particular, we shall examine five different models and for each determine power curves associated with tests of the hypothesis $H_0 : \gamma = 0$ vs $H_A : \gamma > 0$. Each model can be expressed in the form

$$\begin{aligned} R_i|\, b_i &\sim \text{Bin}(n_i, p_i), \quad i = 1, ..., 50, \\ \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_i + f(z_i;\gamma)\cdot b_i, \\ b_i &\sim N(0, \sigma^2). \end{aligned} \tag{5.5.1}$$

We consider the following five formulations:

$M1$: $f(z_i;\gamma) = (\frac{10}{n_i})^\gamma$, $(z_i = n_i)$, $x_i = \begin{cases} 1 & i = 1, ..., 25 \\ 0 & i = 26, ..., 50 \end{cases}$,

$(n_1, n_2, ..., n_{25}) = (20, 40, 60, 80, ..., 500)$,

$n_{i+25} = n_i, \quad i = 1, ..., 25$,

$\beta_0 = -1, \ \beta_1 = 2, \ \sigma^2 = 4$;

$M2$: as $M1$, but $(n_1, n_2, ..., n_{25}) = (20, 28, 35, 42, 50, ..., 200)$ (cluster sizes increasing linearly from 20 to 200, with n_i rounded to the nearest integer);

$M3$: as $M1$, but $(n_1, n_2, ..., n_{25}) = (20, 23, 27, 30, 33, ..., 100)$ (cluster sizes increasing linearly from 20 to 100, with n_i

rounded to the nearest integer);

$M4$: $f(z_i;\gamma) = e^{\gamma x_i}$, $(z_i = x_i)$, $x_i \sim \text{Uniform}(0,1)$, $i = 1,...,50$,
$n_i = 40 \quad \forall\, i$, $\beta_0 = -1$, $\beta_1 = 2$, $\sigma^2 = 0.25$;

$M5$: as $M4$, but $\sigma^2 = 1.0$.

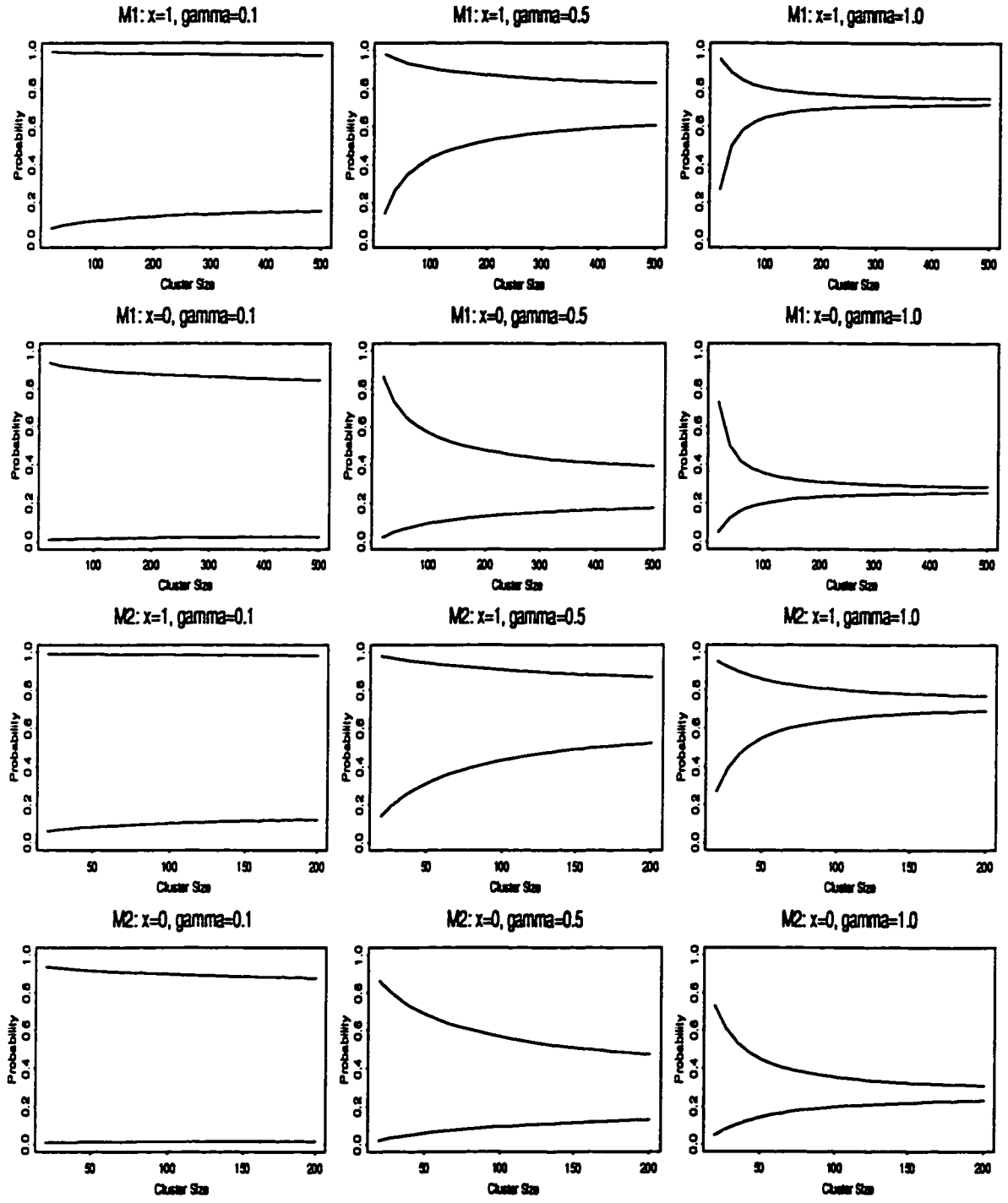
Models $M1$ to $M3$ describe situations in which intra-cluster correlation is a function of the cluster size n_i, for various observed ranges of n_i. In these models the covariate associated with the correlation structure of the data has no impact on the mean. Models $M4$ and $M5$ examine cases where intra-cluster correlation is a function of a continuous covariate, which also appears in the mean formulation of the model.

It is useful to consider what the test $H_0 : \gamma = 0$ vs $H_A : \gamma > 0$ implies about the probabilities p_i, since these are of immediate interest. The specific form of the random component in each cluster should have little bearing on point estimates or predictions one might make from a given model; it impacts rather on the variance associated with an estimated probability. (We consider here only the variability in p_i stemming from the random effect $f(z_i;\gamma)b_i$, not that due to the estimation of β). Figure 5.4 shows upper and lower bounds on the variability of p_i for each of the models $M1$ to $M5$, computed as a function of the covariate z_i in $f(z_i;\gamma)$. The plots may be interpreted as indicating approximate 95 % confidence limits on p_i, since the range at a given set of covariate values was computed as

$$((1 + e^{-(\beta_0+\beta_1 x_i - 2\sigma f(z_i;\gamma))})^{-1}, (1 + e^{-(\beta_0+\beta_1 x_i + 2\sigma f(z_i;\gamma))})^{-1}).$$

If $f(z_i;\gamma)$ is a function of a covariate which has no effect on the mean specification of the model, as in models $M1$ to $M3$, testing $H_0 : \gamma = 0$ is equivalent to testing for parallel lines on a plot of this type (assuming that $\gamma = 0 \rightarrow f(z_i;\gamma) = 1$). If on the other hand z_i

Variance Ranges for p_i, Models $M1$ - $M5$



(continued)

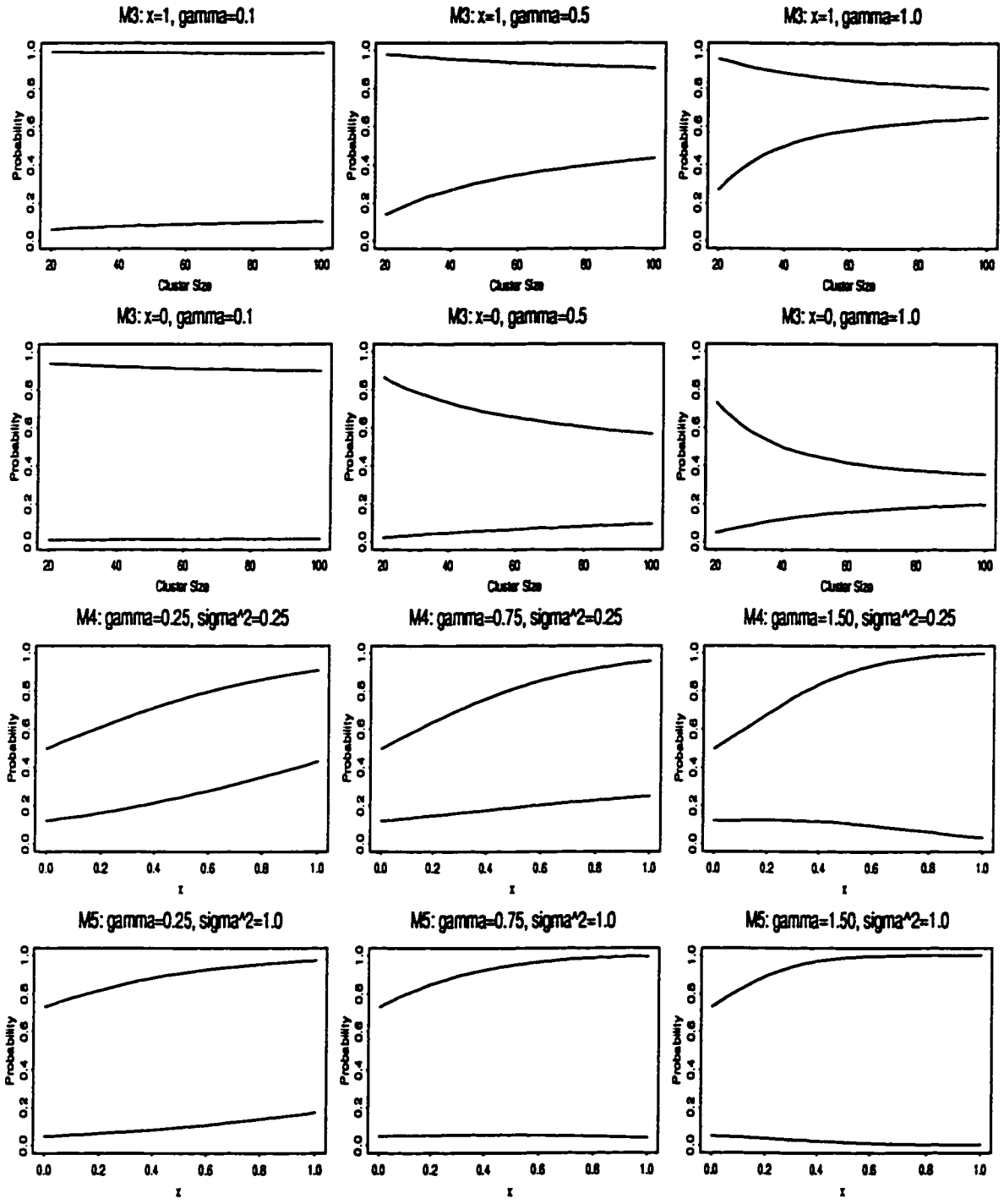


Figure 5.4: Variance ranges for p_i for models $M1$ to $M5$; upper and lower bounds are shown, corresponding to values of $f(z_i;\gamma)b_i$ at a distance of $\pm$ 2 standard deviations from zero.

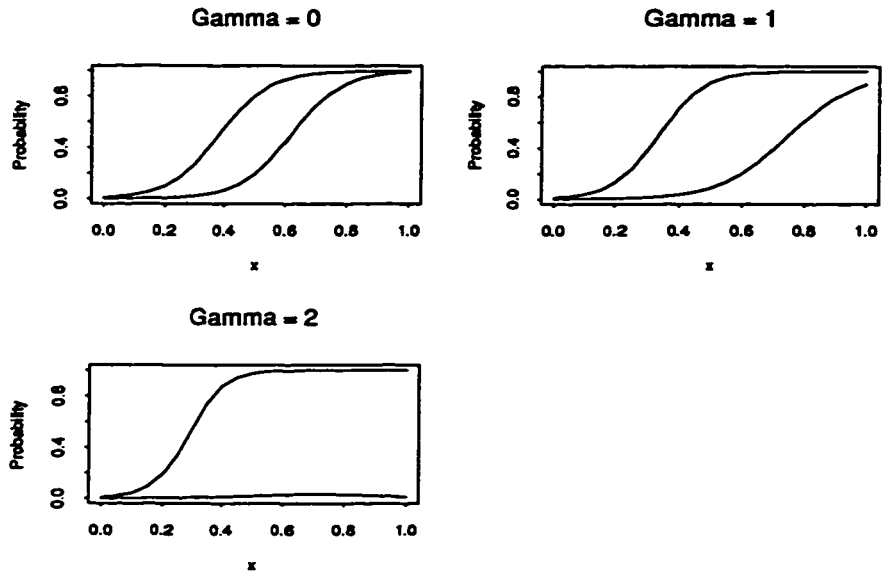


Figure 5.5: Variance ranges for p_i, with $\text{logit}(p_i) = -6 + 12x \pm 2\sigma e^{\gamma x}$, $x \in (0,1)$ and $\sigma^2 = 0.5$.

also appears in the mean formulation of the model, as in $M4$ and $M5$, the null hypothesis is equivalent to two horizontally parallel S-curves, shifted in location by a distance of $4\sigma/\beta$, where β is the coefficient of z_i in the model for the mean; (this is assuming that other predictors for the mean are held constant). It is obvious that as the magnitude of γ increases, the departure from parallel lines or S-curves becomes more dramatic. It is however more difficult to judge whether two curves differ only by a shift in location than it is to determine whether two lines are parallel. An example of the ideal pattern one might detect in the former case, both under H_0 and H_A, is shown in figure 5.5. The three plots show probability ranges

$$((1+e^{-(-6+12x-2\sigma e^{\gamma x})})^{-1}, (1+e^{-(-6+12x+2\sigma e^{\gamma x})})^{-1})$$

for $x \in (0,1)$, $\sigma^2 = 0.5$ and $\gamma = 0, 1$ and 2.

Returning to the task of power calculation, for $M1$ to $M3$, 400 data sets were simulated from each model, assuming each of the following values of γ: $\gamma = 0.0, 0.1, 0.2, 0.3, 0.5, 1.0$. In each case the model was fit to the 400 simulated data sets, and the empirical distribution of the test statistic $\hat{\gamma}/\text{s.e.}(\hat{\gamma})$ calculated. For a given type I error rate α, an estimate of the power for testing $H_0 : \gamma = 0$ vs $H_A : \gamma > 0$ was then computed as the proportion of times an observed value $\hat{\gamma}_r/\text{s.e.}(\hat{\gamma}_r)$, $r = 1, ..., 400$, exceeded the critical value $Z_{1-\alpha}$ from the standard normal distribution. Note that when $\gamma = 0$, the probability of rejecting H_0 reflects a Type I error rate, whereas for $\gamma > 0$, the alternative hypothesis is true and hence the probability of rejecting H_0 reflects the power against the alternative. We proceeded similarly for models $M4$ and $M5$, with the exception that data were simulated for $\gamma = 0.0, 0.25, 0.5, 0.75, 1.0$ and 1.5. Results were obtained for $\alpha = 0.005, 0.01, 0.025, 0.05$ and 0.10, and are shown in figure 5.7. (The power for values of γ not explicitly examined was estimated using cubic spline interpolation). For each model we assessed the assumption of normality of the test statistics by constructing normal probability plots for $\hat{\gamma}_r/\text{s.e.}(\hat{\gamma}_r)$, $r = 1, ..., 400$ under the null hypothesis $H_0 : \gamma = 0$, and by examining histograms of these standardized estimates. The normal plots all produced a linear pattern, roughly about the 45 degree line $y = x$, and the histograms were all symmetric in shape (see figure 5.6), suggesting that it is reasonable to treat the observed test statistics under H_0 as a sample from a standard normal distribution. We also compared the observed proportion of test statistics

	$M1$	$M2$	$M3$	$M4$	$M5$
$\alpha = 0.005$	0.0100	0.0075	0.0075	0.0025	0.0000
$= 0.010$	0.0150	0.0075	0.0125	0.0050	0.0075
$= 0.025$	0.0250	0.0300	0.0375	0.0300	0.0250
$= 0.050$	0.0525	0.0675	0.0450	0.0750	0.0550
$= 0.100$	0.1000	0.1100	0.1000	0.1300	0.1000

Table 5.3: Comparing empirically determined test sizes with α, for models $M1$ to $M5$; entries in the table give the proportion of test statistics greater than $Z_{1-\alpha}$.

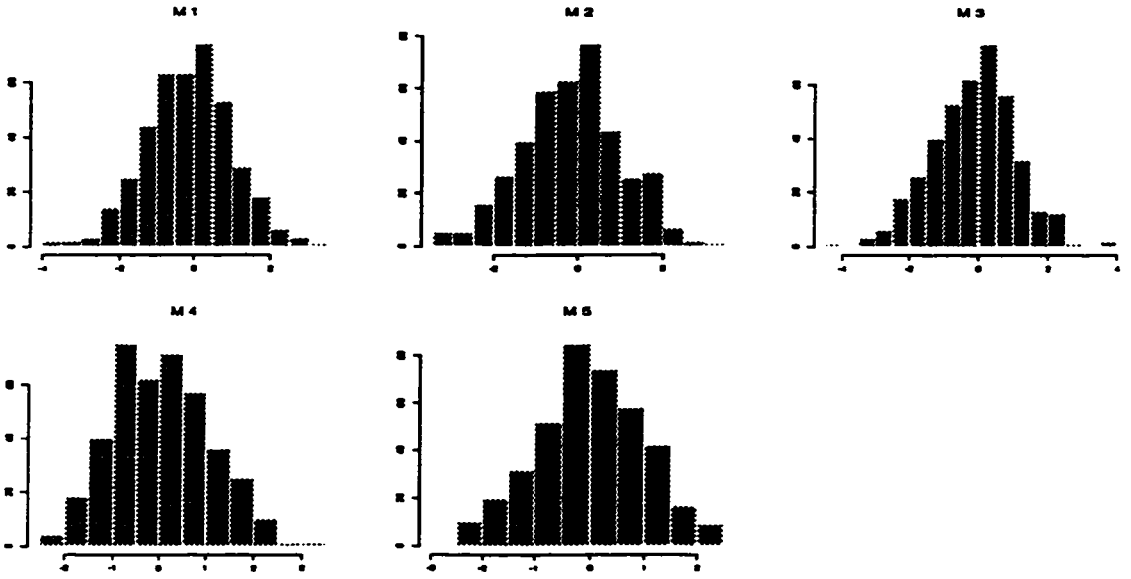


Figure 5.6: Histograms of test statistics $\hat{\gamma}/\text{s.e.}(\hat{\gamma})$ under $H_0 : \gamma = 0$, models $M1$ to $M5$.

greater than $Z_{1-\alpha}$ under H_0 to α, for each of the various sizes of tests considered. Table 5.3 compares the empirically determined test sizes to the true values of α; there appears to be very good agreement between the two, especially considering the moderate number of simulations.

Referring to figure 5.7 we note that for models $M1$ to $M3$, the power to detect a positive value of γ decreases as the range of cluster sizes becomes smaller; intuitively we should expect the power for testing whether or not $\gamma > 0$ to be a function of the effective range of the variance of the random component in the model. In the first three models this range is given by $((10/n_L)^{2\gamma}\sigma^2, (10/n_S)^{2\gamma}\sigma^2)$, where n_S and n_L refer to the smallest and largest cluster size, respectively. Note that this range decreases to 0 as $\gamma \to 0$. Furthermore, it seems that the power depends not so much on the absolute range of the random effects variance (designating $f_i b_i$ to be the random effects in this case) as on the ratio of the upper to the lower endpoint. This is illustrated well in models $M4$ and $M5$, which have very similar power curves; the model specifications are the same with the exception that the

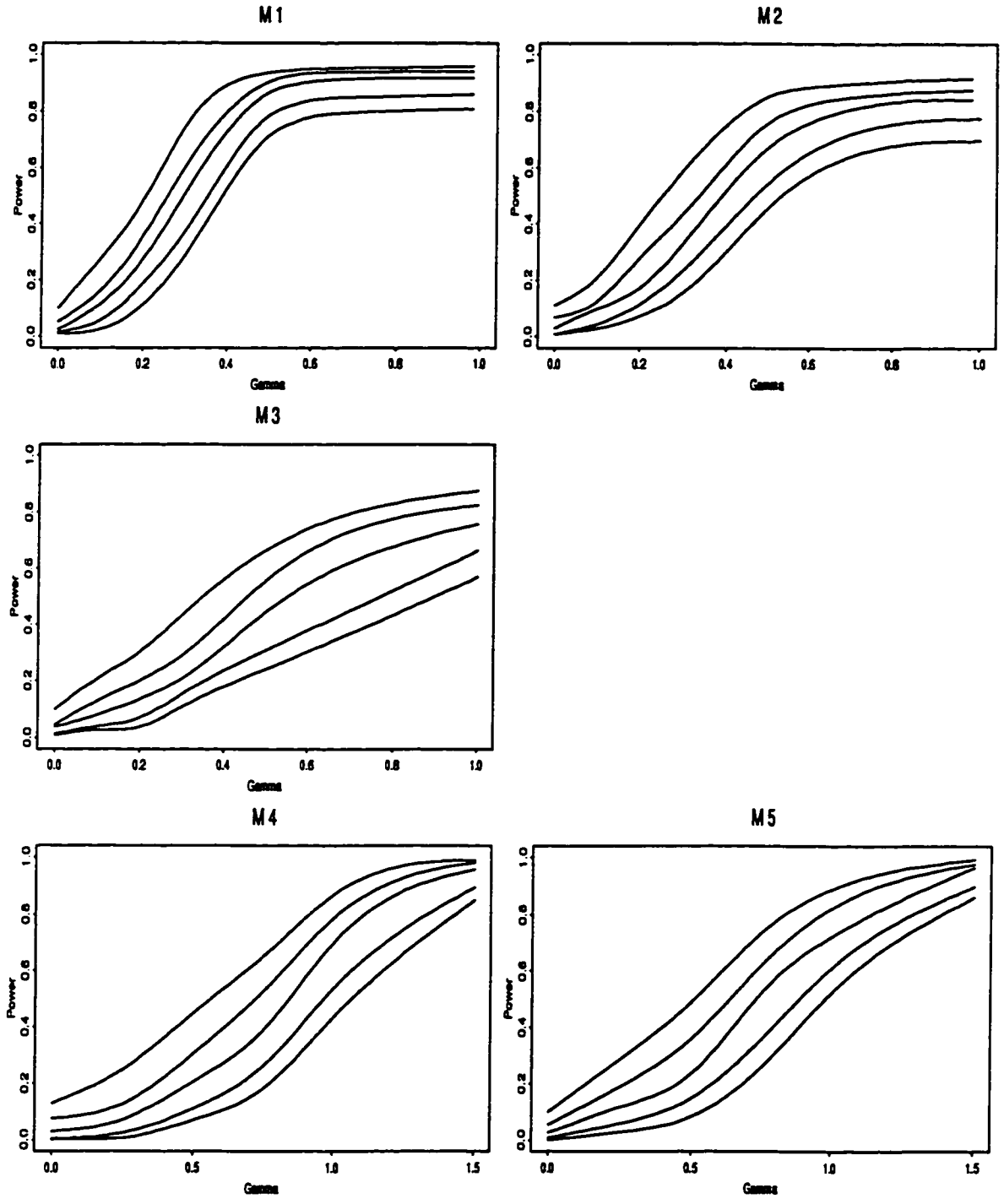


Figure 5.7: Simulated power curves for models $M1$ to $M5$, for the test $H_0 : \gamma = 0$ vs $H_A : \gamma > 0$. Results are shown for 5 type I error rates: 0.005, 0.01, 0.025, 0.05 and 0.10 (corresponding to the curves in order from bottom to top in each plot).

constant prior variance σ^2 in $M5$ is four times greater than in $M4$. Hence the cluster-level variance range, $(\sigma^2, e^{2\gamma}\sigma^2)$, is much larger in absolute terms for $M5$, but the same on a relative scale. Of course one must bear in mind that in testing $H_0 : \gamma = 0$ vs $H_A : \gamma > 0$ we are trying to determine whether the extraneous variation in the data is of a particular form. If σ^2 is very small such a test will be neither powerful nor relevant. In this case the more pertinent question concerns whether or not there is evidence for overdispersion at all.

5.5.2 The Effect of Model Misspecification

We now examine the effect of model misspecification for various cases. As in the previous section we consider a number of model formulations. The first four of these can be expressed as in (5.5.1) and are as follows:

$M6$: $f(z_i;\gamma) = \sqrt{\frac{20}{n_i}}^{\gamma}$, $(z_i = n_i)$, $x_i = \begin{cases} 1 & i = 1,...,25 \\ 0 & i = 26,...,50 \end{cases}$,
$(n_1, n_2, ..., n_{25}) = (20, 40, 60, 80, ..., 500)$ $n_{i+25} = n_i$, $i = 1, ..., 25$,
$\beta_0 = -2$, $\beta_1 = 1$, $\gamma = 1$, $\sigma^2 = 1$;

$M7$: as $M6$, but $(n_1, n_2, ..., n_{25}) = (20, 23, 26, 30, 34, ..., 382, 437, 500)$
(cluster sizes increasing exponentially, so that $\{\log(n_1), ..., \log(n_{25})\}$
is a linearly increasing series (values of n_i rounded to nearest integer));

$M8$: $f(z_i;\gamma) = x_i^{\gamma}$, $(z_i = x_i)$, $x_i \sim \text{Uniform}(0,1)$, $i = 1, ..., 50$,
$n_i = 40 \quad \forall\, i$, $\beta_0 = -1$, $\beta_1 = 2$, $\gamma = 1$, $\sigma^2 = 1$;

$M9$: $f(z_i;\gamma) = z_i^{\gamma}$, $z_i \sim \text{Unif.}(0,1)$, $x_i = \begin{cases} 1 & i = 1,...,25 \\ 0 & i = 26,...,50 \end{cases}$,
$n_i = 40 \quad \forall\, i$, $\beta_0 = -1$, $\beta_1 = 2$, $\gamma = 1$, $\sigma^2 = 1$.

Models $M6$ and $M7$ are similar to $M1$ and $M2$ and examine situations where intra-cluster correlation is a function of cluster size, where either there is a uniform representation of cluster sizes from smallest to largest, or where there are many small to moderately sized clusters, and only a few large ones. $M8$ and $M9$ resemble $M3$ and $M4$ in describing cases in which intra-cluster correlation is a function of a continuous covariate, which may or may not have an effect on the mean of the model.

The last model ($M10$) includes an individual-level covariate; we assume that individuals belong to one of two groups within each cluster and that both the mean and the impact of the cluster-level random effect is different in the two groups. Thus define the model as

$$
\begin{aligned}
M10: \qquad Y_{ij}|\, b_i &\sim \text{Bin}(1, p_{ij}), \quad j = 1, ..., 40, \quad i = 1, ..., 30, \\
\log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \beta_0 + \beta_1 x_{ij} + e^{\gamma x_{ij}} b_i, \\
x_{ij} &= \begin{cases} 1 & j = 1, .., 20 \\ 0 & j = 21, ..., 40 \end{cases}, \\
& \beta_0 = -2, \quad \beta_1 = 1, \quad \gamma = 1, \quad b_i \sim N(0, 0.25).
\end{aligned}
$$

We simulated 300 data sets under each of these five models, and for every data set we fit the true model, as well as the incorrect simpler model, assuming $f(z_{ij}; \gamma) = 1$. The results are summarized in table 5.4. The mean of the estimated model parameters from the 300 data sets is reported, as well as the average of the model-based standard errors (s.e.(M)) and, for comparison, the empirical sample standard errors of the estimators of the parameters (s.e.(S)). In addition, rows indicated by $f_L^2\sigma^2$ and $f_U^2\sigma^2$ report the lower and upper endpoints of the range of the variance component $f^2(z_{ij}; \gamma)\sigma^2$, as well as the estimated endpoints from the correct and misspecified model fits. Thus for example, for $M6$ the range is $(0.04, 1.0)$ and estimated from the simulation to be the average value of $(\sqrt{20/500}^{\,2\hat{\gamma}_r}\hat{\sigma}^2_r, \sqrt{20/20}^{\,2\hat{\gamma}_r}\hat{\sigma}^2_r)$, $r = 1, ..., 300$. Finally, for each model the average value

Model		True Value	Correct Model Fit Mean	s.e.(M)	s.e.(S)	Misspecified Model Fit Mean	s.e.(M)	s.e.(S)
$M6$	β_0	-2.0	-1.9959	.0651	.0711	-1.9897	.0716	.0747
	β_1	1.0	0.9975	.0880	.0942	0.9978	.0971	.1051
	γ	1.0	0.9635	.4004	.5729			
	$f_L^2\sigma^2$	0.04	0.0387			0.0811		
	$f_U^2\sigma^2$	1.00	1.3760			0.0811		
	llik		-6184.14			-6188.18		
$M7$	β_0	-2.0	-2.0054	.0853	.0895	-1.9837	.1055	.1073
	β_1	1.0	1.0057	.1152	.1216	0.9983	.1431	.1492
	γ	1.0	1.0562	.3239	.3590			
	$f_L^2\sigma^2$	0.04	0.0383			0.1753		
	$f_U^2\sigma^2$	1.00	1.1445			0.1753		
	llik		-3653.72			-3659.44		
$M8$	β_0	-1.0	-0.9972	.1146	.1116	-1.0000	.1739	.1296
	β_1	2.0	1.9991	.2937	.2933	1.9627	.2986	.3290
	γ	1.0	1.2214	.5261	.6076			
	$f_L^2\sigma^2$	0.0	0.0000			0.2967		
	$f_U^2\sigma^2$	1.0	1.1395			0.2967		
	llik		-1271.56			-1278.48		
$M9$	β_0	-1.0	-1.0138	.1069	.1100	-1.0132	.1322	.1431
	β_1	2.0	2.0196	.1519	.1580	2.0197	.1875	.2061
	γ	1.0	1.2195	.5813	.6755			
	$f_L^2\sigma^2$	0.0	0.0000			0.3050		
	$f_U^2\sigma^2$	1.0	1.1465			0.3050		
	llik		-1148.86			-1154.91		
$M10$	β_0	-2.0	-2.0274	.1653	.1705	-2.1896	.2381	.1843
	β_1	1.0	1.0195	.3209	.2492	1.2369	.3196	.2245
	γ	1.0	1.2108	.5764	.6393			
	$f_L^2\sigma^2$	0.2500	0.2546			1.0863		
	$f_U^2\sigma^2$	1.8473	1.8580			1.0863		
	llik		-550.28			-554.63		

Table 5.4: Summary of correct and misspecified model fits, models $M6$ to $M10$.

of the maximized log-likelihood is also given.

For $M6$ through $M9$, misspecification of the model does not seem to have a substantial impact on the fixed effect estimates $\hat{\beta}$, nor on their standard errors; if these are the sole quantities of interest no loss is incurred by simply fitting a standard random effects model. The correct model does however provide a better fit to the data in that it gives a better description of the correlation structure. In this case this is also reflected in a significant increase in log-likelihood over the maximized value from fitting the incorrect model. Chapter 6 will discuss the idea of goodness-of-fit of models for correlated data in greater detail.

Interestingly, for model $M10$ erroneously assuming that $f(z_{ij};\gamma) = 1$ appears to produce biased estimates of β_0 and β_1. In the previous models all covariates affecting the mean response and the random effect were on the level of the cluster, and modelling the correlation structure via $f(z_i;\gamma)$ had little impact on the fixed effects parameters specifying the mean, even in model $M8$ where z_i also appeared in the mean formulation. In $M10$ however, not only does the covariate x_{ij} divide the observations in each cluster into two groups, for which we want to estimate separate means, but the group with the larger mean also has much greater variability. None of the other models displays this type of imbalance, which likely explains the results. In cases such as this some reasonable form of modelling the correlation structure seems particularly important.

Overall, with perhaps the exception of model $M10$, the overdispersion in the models above is not extreme, so one might have expected the effect of misspecifying the nature of the extraneous variation to be relatively inconsequential, as far as parameter estimates and their standard errors are concerned. But this fact itself is noteworthy, since it is of interest to consider the effect of misspecifying the correlation structure in cases where it is not immediately obvious. When the impact of the function $f(z;\gamma)$ is very pronounced it is not likely to escape notice. In such situations it is also more likely that model misspecification in the sense discussed here will have a greater effect on parameter estimates and standard

errors. This, however, warrants further investigation.

5.6 Discussion

In this chapter we examined the use of random effects models to approximate correlation structures in clustered binary data. We discussed the relationship between cluster-specific and population-averaged model formulations, pointing out that each type of model can induce a certain form of the other. There is an abundant literature dealing either directly or indirectly with the link between these two types of models. We stress that due to this link it is possible and in many cases easier, at least for cluster-correlated data, to achieve the same end through random effects models as is achieved by a direct modelling of intra-cluster correlation, as in approaches like GEE or the representation given in Bahadur (1961).

The random effects models presented here are ideally suited to drawing mixed-effects model inferences from cluster-correlated data, when specific hypotheses about the correlation structure are of interest, or when information about the correlation structure is available. The WSPP3 elementary intervention, for instance, was based on a social influences curriculum which increased students' awareness of various influences in their environment which might prompt smoking, such as peer pressure, and provided tools to help them resist these pressures. The nature of this intervention program underscores the importance of understanding behavioural patterns in the study population when examining an outcome such as smoking status. In mathematical terms this can be quantified in terms of a non-independent covariance structure, using covariates to determine a specific model. For example, an indicator of individual-level risk (high vs moderate or low) in the random component of a model of form (5.2.1) shows that school-to-school variability is much higher for high-risk students. This suggests that students whose individual profile puts them at high risk of smoking tend to be more similar in their smoking behaviour within a given

school than students at low or moderate risk; see section 7.1.2. Proceeding in a similar fashion, no difference in terms of the correlation structure was observed on the basis of gender.

Model (5.2.1) is useful in that it provides a straightforward tool with which aspects of the correlation structure of the data can be captured in a parsimonious fashion. Instead of resorting to several variance components to explain the extraneous variability, covariate information (z) is used in conjunction with associated parameters (γ) to adjust the impact of random effects from a single univariate distribution. As pointed out in section 5.2, added flexibility is gained by the fact that γ, which determines the impact of z, is estimated from the data.

As indicated earlier, the work of Neuhaus et al. (1992) has special relevance to this discussion. The authors examine the effects of mixture distribution misspecification in mixed-effects logistic models. Their findings indicate that estimated regression coefficients from the fit of a misspecified model will be nearly consistent; the intercept term will also show little bias if both the true and assumed random effects distributions are symmetric, but larger bias if the true mixing distribution is skewed and the assumed one is symmetric. Furthermore, simulations similar to those carried out in section 5.5.2 revealed that valid standard error estimates for the regression parameters can also be obtained from the misspecified model. Our work centers more on incorporating covariate information into the random effects specification of the mixed model. We have considered the effects of misspecification in the sense of omitting such covariates when the random component of the model truly depends on them. Throughout we have assumed that the underlying mixing distribution is normal. Future research should combine the focusses of Neuhaus et al. (1992) and the work presented here in studying the impact of model misspecification when the mixing distribution depends on covariates and is not normal or symmetric. Situations in which the same covariate(s) affect both the mean and the correlation structure of the responses are of special interest, in particular cases such as model $M10$ (section 5.5.2),

involving imbalances between the mean and the correlation structure.

The simulation results described in section 5.5, while limited, give a general indication of the behaviour of the models presented here. It would be useful to undertake further study of these models through more extensive simulations.

Chapter 6

Testing the Goodness-of-Fit of Models for Correlated Data

6.1 Introduction

In the previous chapters we were primarily concerned with issues surrounding model specification and estimation. However a third important component of the model fitting process, and one which does not always receive due consideration, is that of checking the adequacy of the fit of the model to the data. Whereas a great deal of research in recent years has focussed on parameter estimation in generalized linear mixed models and other extensions of GLMs for non-independent data, comparatively little attention has been directed toward the problem of assessing model fit. This is due in part to the fact that in relaxing the assumption of independence between observations, the likelihood function usually becomes much more complicated, and when recourse is taken to other methods of analysis the notion of a clearly defined model deviance is no longer available. The deviance, however, is often the measure on which goodness-of-fit statistics are based, at least for nested model comparisons. The problem of assessing goodness-of-fit for correlated data is further

complicated since one needs to entertain a broader concept of the idea of 'fit': this must refer not only to how well fitted values match the observed data, but also to how well the dispersion predicted by the model matches the observed variance in the data, given a certain specification for the mean response.

Though not abundant, the literature in this area does include several noteworthy references. Vonesh, Chinchilli and Pu (1996) describe a general, formal approach to assessing the adequacy of an assumed mean and covariance structure in the framework of generalized nonlinear mixed-effects models. The authors propose a statistic similar to the R^2 criterion for linear regression models, for testing the fit of the predicted values from a model, as well as a pseudo-likelihood ratio test for testing the adequacy of the assumed covariance structure. The principal drawback is the fact that both procedures are based on the underlying assumption of normality and are therefore not well suited to problems involving discrete data. For binary data, methods for testing goodness-of-fit have been proposed based on partitioning the covariate space in some meaningful manner and considering functions of the observed minus predicted responses in each region. See for example Tsiatis (1980) and Hosmer and Lemeshow (1989). Extensions of this idea to correlated data are possible by considering differences between data-based and fully model-based variance contributions in each region, where these terms are subject to some definition; examples of this will be seen in sections 6.3 and 6.4. Lipsitz, Fitmaurice and Molenberghs (1996) consider partitioning response categories to construct goodness-of-fit statistics for ordinal response models. Other recent references in general include le Cessie and van Houwelingen (1995), Farrington (1996) and section 3.3 in Ng (1997).

In the following section we propose a procedure to assess the fit of a model to non-independent data, under very general conditions. It is quite straightforward, yet nevertheless appears to give valuable insight into how well a given model reflects the covariance structure of the data, having assumed that a given mean specification cannot be improved upon. We propose both numerical statistics as well as associated plots to assess goodness-

of-fit.

The developments in this chapter will pertain to cases in which there is a single source of clustering, so that observations are correlated within but independent between clusters.

6.2 Covariance-based Measures of Goodness-of-Fit

6.2.1 Motivation

Any definition of goodness-of-fit involves some notion of the extent to which predicted values from a model ($\hat{\boldsymbol{Y}}$) agree with the data observed ($\boldsymbol{Y}$). Assuming that the form of the chosen model is correct, improvements in fit are then attempted by refining the formulation for $\mathrm{E}(\boldsymbol{Y})$, and evaluated with a test for goodness-of-fit. With clustered data, however, one is often faced with overdispersion of some form, even when the mean structure of the data has been modelled as carefully as possible. In such cases the resulting overdispersion can only be corrected by a more careful modelling of the covariance structure. As an interesting example, consider the data from a toxicology experiment cited in Ganio and Schafer (1992). This study was conducted to investigate the carcinogenic effects of the toxic compound aflatoxin. Forty tanks of rainbow trout embryos, each containing between 84 and 90 fish, were exposed to either aflatoxin B1 (AB1) or aflatoxicol (Al), a related substance. Exposure occurred for one hour at one of 5 doses: 0.01, 0.025, 0.05, 0.10 or 0.25 parts per million; recorded was the incidence of liver tumors after one year, in each of the tanks. A logistic model adjusting for the effects of dose and compound type, as well as their interaction, seems to fit quite well at first glance, but examining the individual deviance components actually suggests underdispersion among the responses from the AB1 tanks, and overdispersion in those from the Al tanks. Hence a truly well-fitting model for these data must have a covariance structure sufficiently flexible to accomodate this phenomenon, and it would be desirable to have a way of assessing whether a given covariance structure

improves on a simpler alternative, in the same manner that other goodness-of-fit measures evaluate the improvement on a given mean response function (see section 6.4 for a detailed discussion of this example). As stated by Vonesh et al. (1996), p.577, "the utility of any statistical model is determined not only by its ability to approximate an otherwise unknown response function but also by its ability to account for the accompanying background noise". In a similar spirit as these authors, we shall therefore be concerned with assessing how well the covariance structure of a given model fits the empirical covariance structure as reflected by the data, assuming at the outset that the mean response function has been correctly specified.

6.2.2 Analysis of $\mathrm{Var}(\sum_{ij} Y_{ij})$

Let Y_{ij} denote the response of individual j in cluster i, $j = 1, ..., n_i$, $i = 1, ..., K$, assuming independence between observations from different clusters. Let $\mathrm{E}(Y_{ij}) = p_{ij}$, and $\boldsymbol{Y}$ be the entire data vector, with $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{p}$. Further, let the parameter vector $\boldsymbol{\rho}$ denote the non-zero intra-cluster correlations. We base our measure of the empirical variance in the data on a quantity $V_D(\boldsymbol{Y}; \boldsymbol{p})$, which we call the data-based variance of $\sum_{ij} Y_{ij}$ and define as

$$\begin{aligned} V_D(\boldsymbol{Y}; \boldsymbol{p}) &= \sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{ij} - p_{ij})^2 + 2 \sum_{i=1}^{K} \sum_{j<j'} (Y_{ij} - p_{ij})(Y_{ij'} - p_{ij'}) \\ &= A_D(\boldsymbol{Y}; \boldsymbol{p}) + B_D(\boldsymbol{Y}; \boldsymbol{p}). \end{aligned} \tag{6.2.1}$$

Define $V_M(\boldsymbol{p}, \boldsymbol{\rho})$, the model-based variance of $\sum_{ij} Y_{ij}$, as

$$\begin{aligned} V_M(\boldsymbol{p}, \boldsymbol{\rho}) &= \sum_{i=1}^{K} \sum_{j=1}^{n_i} \mathrm{Var}(Y_{ij}) + 2 \sum_{i=1}^{K} \sum_{j<j'} \mathrm{Cov}(Y_{ij}, Y_{ij'}) \\ &= A_M(\boldsymbol{p}, \boldsymbol{\rho}) + B_M(\boldsymbol{p}, \boldsymbol{\rho}), \end{aligned} \tag{6.2.2}$$

where $\text{Var}(Y_{ij})$ and $\text{Cov}(Y_{ij}, Y_{ij'})$ depend on a given model specification. We wish to compare $V_D(\boldsymbol{Y}; \boldsymbol{p})$ with $V_M(\boldsymbol{p}, \boldsymbol{\rho})$. Note that under the true model for the data,

$$V_M(\boldsymbol{p}, \boldsymbol{\rho}) \;=\; \text{E}[V_D(\boldsymbol{Y}; \boldsymbol{p})] \;=\; \text{Var}(\sum_{ij} Y_{ij}).$$

In practice we need to estimate $V_D(\boldsymbol{Y}; \boldsymbol{p})$ and $V_M(\boldsymbol{p}, \boldsymbol{\rho})$ by replacing $\boldsymbol{p}$ and $\boldsymbol{\rho}$ with their estimated values from a given model under consideration. (Note that although largely empirical, $V_D(\boldsymbol{Y}; \boldsymbol{p})$ is also weakly model-dependent in that it depends on the mean formulation). Thus we may write

$$\begin{aligned} \hat{V}_D \;&=\; V_D(\boldsymbol{Y}; \hat{\boldsymbol{p}}) \;=\; \hat{A}_D + \hat{B}_D \qquad \text{and} \\ \hat{V}_M \;&=\; V_M(\hat{\boldsymbol{p}}, \hat{\boldsymbol{\rho}}) \;=\; \hat{A}_M + \hat{B}_M. \end{aligned}$$

(Note that to compute $\hat{V}_M$ we need estimates of the marginal covariances between Y_{ij} and $Y_{ij'}$. These are directly available for marginal models, but usually have to be approximated for conditional models. For this latter class of models, in fact, some adjustment is generally also necessary to obtain the marginal means. See section 6.3 for further comments on this). The goal is to find a model which fits well, in the sense that $|\hat{V}_D - \hat{V}_M|$ is small.

To determine whether or not $\hat{V}_D$ and $\hat{V}_M$ differ significantly in a statistical sense, we need to have an idea of the sampling variability in $\hat{V}_D$. If covariates are on the level of the cluster only, so that $p_{ij} = p_i$ for all j and $\text{Corr}(Y_{ij}, Y_{ij'}) = \rho_i$, $j \neq j'$, it is possible and feasible to compute $\text{Var}(A_D(\boldsymbol{Y}; \boldsymbol{p}))$ and $\text{Var}(B_D(\boldsymbol{Y}; \boldsymbol{p}))$ exactly. Straightforward calculations show that

$$\begin{aligned} \text{Var}(A_D(\boldsymbol{Y}; \boldsymbol{p})) \;&=\; \text{Var}\{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{ij} - p_{ij})^2\} \\ &=\; \sum_{i=1}^{K} n_i p_i (1 - p_i)(1 - 2p_i)^2 [1 + \rho_i (n_i - 1)]. \qquad (6.2.3) \end{aligned}$$

Some further algebra yields the result

$$\begin{aligned}\mathrm{Var}(B_D(\boldsymbol{Y};\boldsymbol{p})) &= \mathrm{Var}\{2\sum_{i=1}^{K}\sum_{j<j'}(Y_{ij}-p_{ij})(Y_{ij'}-p_{ij'})\} \\ &= 4\sum_{i=1}^{K}\{\binom{n_i}{2}[\rho_i p_i(1-p_i)\{1-\rho_i p_i(1-p_i)\}+ \\ &\qquad p_i^2(1-p_i)^2\{1-4\rho_i\}]+2\Psi_i\}, \qquad (6.2.4)\end{aligned}$$

where

$$\Psi_i = 3\binom{n_i}{4}\{\mathrm{E}(Y_{ij}Y_{il}Y_{im}Y_{in})-4p_i\mathrm{E}(Y_{ij}Y_{il}Y_{im})+3p_i^4+6\rho_i p_i^3(1-p_i)-\rho_i^2 p_i^2(1-p_i)^2\}+$$
$$3\binom{n_i}{3}\{(1-2p_i)\mathrm{E}(Y_{ij}Y_{il}Y_{im})+5\rho_i p_i^3(1-p_i)-2\rho_i p_i^2(1-p_i)-\rho_i^2 p_i^2(1-p_i)^2-p_i^3+2p_i^4\},$$
$$j\neq l\neq m\neq n.$$

We shall estimate $\mathrm{Var}(A_D(\boldsymbol{Y};\hat{\boldsymbol{p}}))$ and $\mathrm{Var}(B_D(\boldsymbol{Y};\hat{\boldsymbol{p}}))$ by replacing p_i and ρ_i with $\hat{p}_i$ and $\hat{\rho}_i$ in (6.2.3) and (6.2.4). Now in order to compute Ψ_i we need to calculate the fourth and third order mean products $\mathrm{E}(Y_{ij}Y_{il}Y_{im}Y_{in})$ and $\mathrm{E}(Y_{ij}Y_{il}Y_{im})$. If $p_{ij}=p_i$ for all j, $\mathrm{E}(Y_{ij}Y_{il}Y_{im}Y_{in})$ takes the same value for all $j\neq l\neq m\neq n$. Further, if $\rho_i>0$, the beta-binomial distribution may be used to compute these mean products. Letting R be distributed as beta-binomial(a_i,b_i) with $n_i=4$, where $a_i=p_i(\frac{1}{\rho_i}-1)$ and $b_i=(1-p_i)(\frac{1}{\rho_i}-1)$, we have in this case that

$$\begin{aligned}E_{4i} = \mathrm{E}(Y_{ij}Y_{il}Y_{im}Y_{in}) &= P(R=4) \\ &= \frac{a_i(a_i+1)(a_i+2)(a_i+3)}{(a_i+b_i)(a_i+b_i+1)(a_i+b_i+2)(a_i+b_i+3)}.\end{aligned}$$

Similarly,

$$E_{3i}=\mathrm{E}(Y_{ij}Y_{il}Y_{im})=\frac{a_i(a_i+1)(a_i+2)}{(a_i+b_i)(a_i+b_i+1)(a_i+b_i+2)}.$$

Note that this parameterization does not allow for negative correlation, since a_i and b_i are

positive parameters in the beta-binomial distribution. If however we express E_{3i} and E_{4i} as functions of p_i and ρ_i, we get

$$E_{3i} = \frac{\prod_{k=0}^{2}\{p_i[\rho_i^{-1}-1]+k\}}{\prod_{k=0}^{2}\{\rho_i^{-1}-1+k\}} \quad \text{and} \quad E_{4i} = \frac{\prod_{k=0}^{3}\{p_i[\rho_i^{-1}-1]+k\}}{\prod_{k=0}^{3}\{\rho_i^{-1}-1+k\}}, \tag{6.2.5}$$

which coincides with the alternative parameterization above when $\rho_i > 0$, but is also well-defined for $\rho_i \leq 0$. The values in (6.2.5) in fact correspond to those obtained from the extended beta-binomial model, described in Prentice (1986), which allows for negative correlation. We therefore compute E_{3i} and E_{4i} using (6.2.5) in general, ensuring of course that ρ_i lies in the admissible range $(-(n_i-1)^{-1}, 1)$.

It would be considerably more cumbersome to derive a similar formula for $\mathrm{Var}(V_D(\boldsymbol{Y};\boldsymbol{p}))$, due to the non-trivial covariance term $\mathrm{Cov}(A_D, B_D)$. Furthermore this is not necessary. Having explicitly expressed V_D and V_M as sums of two terms A and B, we note that A is the contribution to the total variance due to the nominal dispersion, and B is the contribution due to over- (or under-) dispersion. If the mean of the model is correctly specified, from which follows the form of the marginal variances in the exponential family, then $\hat{A}_D$ and $\hat{A}_M$ are generally very similar, and the comparison one is really interested in is $\hat{B}_D - \hat{B}_M$. To illustrate, we briefly consider the analysis of the toxicology data introduced in section 6.2.1. Saving the details for section 6.4, three models were fit to these data: a simple logistic model and two GEE models which address the correlation structure in different ways. Table 6.1 reports the values of $\hat{A}_D$ and $\hat{A}_M$, as well as the standard error $\sqrt{\widehat{\mathrm{Var}}(A_D(\boldsymbol{Y};\hat{\boldsymbol{p}}))}$ in brackets for each of these models. The close proximity of $\hat{A}_D$ and $\hat{A}_M$ seen here is typical of what we have observed in other real and simulated data sets and emphasizes that any discrepancy between $\hat{V}_D$ and $\hat{V}_M$ is bound to be due to a large difference between $\hat{B}_D$ and $\hat{B}_M$, indicating an incorrect modelling of the correlation structure. We will therefore focus on $\hat{B}_D$ and $\hat{B}_M$.

Equation (6.2.4) provides an exact expression for $\mathrm{Var}(B_D(\boldsymbol{Y};\boldsymbol{p}))$ when covariates are

	Logistic Model		GEE-M1		GEE-M2	
$\hat{A}_D$	580.72	(11.10)	580.72	(9.73)	580.72	(10.08)
$\hat{A}_M$	580.72		580.80		580.96	

Table 6.1: Comparison of $\hat{A}_D$ and $\hat{A}_M$ for the toxicology data given in Ganio and Schafer (1992).

on the level of the cluster only. Whereas an analogous expression can be developed for the case of individual-level covariates, this becomes computationally intractable even for moderately sized clusters. In addition, an estimate of $\text{Var}(B_D(\boldsymbol{Y};\boldsymbol{p}))$ does not give any additional information beyond the second moment about the distribution of $B_D(\boldsymbol{Y};\boldsymbol{p})$. For these reasons we use simulation as a means for estimating $\text{Var}(\hat{B}_D)$ in cases where an exact result is difficult to compute, and to study the shape of the distribution of $\hat{B}_D$. To this end we shall treat the estimated values $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{\rho}}$ from a given model fit as fixed, and proceed as follows:

1. Fit the model of interest, $\mathcal{M}$, to the original data $\boldsymbol{Y}$ to obtain $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{\rho}}$ (if $\mathcal{M}$ has a non-independent correlation structure).

2. Generate data $\boldsymbol{Y}^*$ from model $\mathcal{M}$, using $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{\rho}}$ (if $\mathcal{M}$ was parameterized in terms of $\boldsymbol{p}$ and $\boldsymbol{\rho}$), or the original parameter estimates from the fit in (1) (if $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{\rho}}$ had to be computed from these).

3. Compute $B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}})$.

4. Repeat steps (2) and (3) N times.

This will give rise to $\hat{B}^*_{D_1}, \hat{B}^*_{D_2}, ..., \hat{B}^*_{D_N}$. Using these values we compute the simulation-based variance estimate

$$\widehat{\text{Var}}(B_D(\boldsymbol{Y};\hat{\boldsymbol{p}})) = \frac{\sum_{s=1}^{N}\{\hat{B}^*_{D_s} - \overline{B^*_D}\}^2}{N-1}, \tag{6.2.6}$$

where $\overline{B}^*_D = \sum_{s=1}^{N} \hat{B}^*_{D,s}/N$.

A model which reflects the correlation structure of the data well should therefore yield a data-based estimate $\hat{B}_D$ reasonably close to $\hat{B}_M$, where what is reasonable is determined in part by the shape of the histogram of the simulated values $\{\hat{B}^*_D\}$. If the degree of skewness is not unduly severe, then $\hat{B}_D$ and $\hat{B}_M$ might be expected to be within two standard errors of one another. In other words the standardized statistic

$$\hat{S}_B = \frac{\hat{B}_D - \hat{B}_M}{\sqrt{\widehat{\text{Var}}(\hat{B}_D)}} \tag{6.2.7}$$

should be in the range (-2, 2).

To investigate how well a given model captures the correlation structure of the data at the cluster level, one can examine relevant partitions of $\hat{B}_D$ and $\hat{B}_M$; this is discussed in the following section.

6.2.3 Partitioning of $\hat{B}_D$, $\hat{B}_M$

As alluded to in section 6.1, the idea of partitioning the covariate space into meaningful groups and comparing the contributions to $\hat{B}_D$ and $\hat{B}_M$ in each group can be useful in detecting somewhat more subtle correlation patterns. This proves to be effective for instance in the toxicology example mentioned earlier, in which tanks should be grouped according to compound type. Alternatively, recall model $M6$, described in section 5.5.2, in which the variance of the random component in the linear predictor for each cluster is a function of cluster size. In comparing the fit of such a model to one which assumes this variance is common across all clusters, one might divide the data into groups of clusters of similar sizes and examine separately the contributions to $\hat{B}_D$ and $\hat{B}_M$ from each group. We shall pursue this particular example further in section 6.3.

In general, suppose a partition of the covariance terms B into m parts is of interest, so

that B can be written as

$$B = B_1 + B_2 + \cdots + B_m.$$

Noting that B is a sum of K independent contributions, one from each cluster, we can also express B as

$$B = \sum_{i=1}^{K} d_i = \sum_{i \in \Omega_1} d_i + \sum_{i \in \Omega_2} d_i + \cdots + \sum_{i \in \Omega_m} d_i,$$

where $\{\Omega_1, ..., \Omega_m\}$ partitions the set of clusters so that $\sum_{i \in \Omega_\ell} d_i = B_\ell$, $\ell = 1, ..., m$. As indicated in the previous section, we can compute data-based as well as model-based estimates of d_i. These are given by

$$\hat{d}_{D_i} = 2 \sum_{j<j'} (Y_{ij} - \hat{p}_{ij})(Y_{ij'} - \hat{p}_{ij'}) \quad \text{and} \quad \hat{d}_{M_i} = 2 \sum_{j<j'} \widehat{\text{Cov}}(Y_{ij}, Y_{ij'}).$$

Hence $\hat{B}_D = \sum_{\ell=1}^{m} \hat{B}_{D_\ell}$ and $\hat{B}_M = \sum_{\ell=1}^{m} \hat{B}_{M_\ell}$, with

$$\hat{B}_{D_\ell} = \sum_{i \in \Omega_\ell} \hat{d}_{D_i} \quad \text{and} \quad \hat{B}_{M_\ell} = \sum_{i \in \Omega_\ell} \hat{d}_{M_i}, \quad \ell = 1, ..., m.$$

Specific questions concerning the fit of a given model, regarding in particular its correlation structure, can be investigated by looking at the standardized values

$$\hat{S}_{B_\ell} = \frac{\hat{B}_{D_\ell} - \hat{B}_{M_\ell}}{\sqrt{\widehat{\text{Var}}(\hat{B}_{D_\ell})}}, \quad \ell = 1, ..., m. \tag{6.2.8}$$

As above, we can estimate $\text{Var}(\hat{B}_{D_\ell})$ using either equation (6.2.4) or a simulation-based estimate. In either case, $\widehat{\text{Var}}(\hat{B}_{D_\ell}) = \sum_{i \in \Omega_\ell} \widehat{\text{Var}}(\hat{d}_{D_i})$, where from (6.2.4),

$$\widehat{\text{Var}}(\hat{d}_{D_i}) = 4\{\binom{n_i}{2}[\hat{\rho}_i\hat{p}_i(1-\hat{p}_i)\{1-\hat{\rho}_i\hat{p}_i(1-\hat{p}_i)\} + \hat{p}_i^2(1-\hat{p}_i)^2\{1-4\hat{\rho}_i\}] + 2\hat{\Psi}_i\}, \tag{6.2.9}$$

or, for the simulation-based estimate, the individual terms in each of $\hat{B}^*_{D_1}, \hat{B}^*_{D_2}, ..., \hat{B}^*_{D_N}$ are

used to compute $\widehat{\text{Var}}(\hat{d}_{D_i})$. That is, noting that $\hat{B}^*_{D_s} = \hat{d}^*_{D_{s1}} + \cdots + \hat{d}^*_{D_{sK}}$, $s = 1, ..., N$, we obtain the simulation based estimates

$$\widehat{\text{Var}}(\hat{d}_{D_i}) = \frac{\sum_{s=1}^{N} \{\hat{d}^*_{D_{si}} - \overline{d}^*_{D_i}\}^2}{N-1}, \qquad i = 1, ..., K, \tag{6.2.10}$$

where $\overline{d}^*_{D_i} = \sum_{s=1}^{N} \hat{d}^*_{D_{si}}/N$.

It is also useful to examine plots of the individual standardized values

$$\hat{S}_i = (\hat{d}_{D_i} - \hat{d}_{M_i})/\sqrt{\widehat{\text{Var}}(\hat{d}_{D_i})}$$

for each cluster. This provides a visual assessment of the model fit and allows one to identify outliers (i.e. clusters inconsistent with the model-based correlation structure). Such plots are considered in section 6.3 and also in the example in section 6.4.

This discussion has only focussed on partitioning the covariate space according to groups of clusters. One might also consider meaningful groupings of the observations within clusters. Recall for example model $M10$, also from section 5.5.2, under which each individual in a given cluster belongs to one of two groups, with the impact of the cluster-level random effect differing in the two groups. Here one might for a given model fit examine separately the contributions to $\hat{B}_D$ and $\hat{B}_M$ (across clusters) from individuals belonging to the same group.

Finally, we have considered only cross-sectionally clustered data here. But the same ideas apply in principle when the dependence between observations is due to longitudinal correlation, though it will be difficult to distinguish minor differences in the intra-individual correlation structure. Also, unless a reasonably large number of repeated observations is available on each subject, plots of the type discussed above will not be very informative.

6.3 Analysis of $|\hat{B}_D - \hat{B}_M|$ Based on Simulated Data

In this section we examine a number of the data sets which were simulated under model $M6$ as described in section 5.5.2. Recall that we specified this model as follows:

$$\begin{aligned} R_i|\, b_i &\sim \text{Bin}(n_i, p_i(b_i)), \quad i = 1, ..., 50, \\ \log\left(\frac{p_i(b_i)}{1-p_i(b_i)}\right) &= \beta_0 + \beta_1 x_i + \sqrt{\frac{20}{n_i}}^{\gamma} b_i, \qquad b_i \sim N(0, \sigma^2), \\ x_i &= \begin{cases} 1 & i = 1, .., 25 \\ 0 & i = 26, ..., 50 \end{cases}, \\ (n_1, n_2, ..., n_{25}) &= (20, 40, 60, ..., 500), \quad n_{i+25} = n_i, \quad i = 1, ..., 25, \end{aligned} \tag{6.3.1}$$

with the true parameter values taken to be $(\beta_0, \beta_1, \gamma, \sigma^2) = (-2, 1, 1, 1)$. For each simulated data set we computed $\hat{\omega}$, the estimated range (over the span of cluster sizes) of the variance of the random term $\sqrt{20/n_i}^{\gamma} b_i$. That is, we calculated $\hat{\omega}_r = \hat{\sigma^2}_r - (20/500)^{\hat{\gamma}_r} \hat{\sigma^2}_r$, $r = 1, ..., 300$, where the subscript r denotes the estimate from the rth simulated data set. We then selected those data sets among the 300 giving the largest and smallest estimated ranges, the endpoints of the interquartile range and the median value. This corresponded to values for $\hat{\gamma}$ and $\hat{\sigma^2}$ giving estimated ranges $\hat{\omega} = 0.001$, 0.392, 0.907, 1.683 and 9.869. We wish to compare the fit of the true model to these data with that of the misspecified model which assumes that $\gamma = 0$ ($\Leftrightarrow \omega = 0$); the idea is to illustrate that the proposed method of assessing model fit will indeed reveal a poorer fit for the misspecified model in cases where there is evidence that the variance range is larger than zero, and an adequate fit in cases where ω seems to be close to zero. We will show that as the magnitude of $\hat{\omega}$ increases, the discrepancy in the fit, in the sense we have discussed, of the true model versus that of the misspecified model increases correspondingly, with only the former continuing to fit well.

In order to compute $\hat{B}_D$ and $\hat{B}_M$ for random effects models such as (6.3.1) we need an estimate $\hat{p}_i$ of the marginal probability $\text{E}_b(p_i(b_i))$, as well as estimates of the marginal

correlations. Using cluster-specific parameter estimates one could simply apply the approximations discussed in section 5.3. Recall however that the formula for ρ_i (equation (5.3.9)) may not be very accurate for large values of σ^2, or $f_i^2\sigma^2$. We therefore recommend using numerical integration for computing the marginal covariance terms for random effects models, as well as the marginal means.

Model 6.3.1 suggests that the cluster-level variability is a decreasing function of cluster size. To investigate how well a given model reflects this characteristic of the data, we shall compute not only $\hat{B}_D$ and $\hat{B}_M$ but examine in particular a breakdown of these terms into four contributions from clusters of similar sizes. The following assignment of clusters was chosen for the four groups:

Ω_1 : contains clusters sized 80 or smaller $(i = 1, ..., 4 \quad \text{and} \quad 26, ..., 29)$

Ω_2 : contains clusters sized 100 - 200 $(i = 5, ..., 10 \quad \text{and} \quad 30, ..., 35)$

Ω_3 : contains clusters sized 220 - 340 $(i = 11, ..., 17 \quad \text{and} \quad 36, ..., 42)$

Ω_4 : contains clusters sized 360 - 500 $(i = 18, ..., 25 \quad \text{and} \quad 43, ..., 50).$

What we should find is that the misspecified random effects model, assuming no dependence of the cluster-level variability on n_i, underestimates the dispersion in smaller clusters and overestimates it in larger ones. For each of the five data sets described above, table 6.4 contains the statistic $\hat{S}_B$ as defined in equation (6.2.7), as well as $\hat{S}_{B_1}, ..., \hat{S}_{B_4}$ (equation (6.2.8)), for the following model fits: the correct model, as in (6.3.1); the misspecified model assuming $\gamma = 0$; and finally, for the sake of comparison, the standard logistic model. In this case maximum likelihood was used to fit all models.

Recalling that $\hat{S}_B$ and $\hat{S}_{B_1}, ..., \hat{S}_{B_4}$ can be computed in two ways, depending on how the relevant standard errors are calculated, we give both estimates for these quantities in table 6.4. Rows flagged by (E) refer to values standardized by variance estimates based

on the exact formula given in the previous section, and those flagged by (S) refer to values standardized by simulation-based variance estimates, computed from $N = 200$ simulated data sets. To generate data $\boldsymbol{Y}^* = (Y_1^*, ..., Y_{50}^*)$ from model (6.3.1), for example, we first obtain a random vector $(b_1^*, ..., b_{50}^*)$ from the $N(0, \hat{\sigma^2})$ distribution. We then compute $p_i(b_i^*) = (1 + \exp\{-(\hat{\beta}_0 + \hat{\beta}_1 x_i + \sqrt{20/n_i}\,\hat{\gamma} b_i^*)\})^{-1}$ and generate an observation Y_i^* $(= R_i^*)$ from the Binomial$(n_i, p_i(b_i^*))$ distribution, for $i = 1, ..., 50$. Here $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}$ and $\hat{\sigma^2}$ refer to the estimates from the fit of model (6.3.1) to the original data.

Examining the histograms of the simulated values $\{\hat{B}_D^*\}$ for these models revealed that the distribution of $\hat{B}_D$ tends to be only moderately right-skewed. One could study this sampling distribution more closely by increasing the number of simulations, and in fact an essentially exact test of the hypothesis $H_0 : \hat{B}_D = \hat{B}_M$ is obtained as $N \to \infty$. However for practical purposes, even reporting the standardized statistics discussed above is sufficient to provide a legitimate quantitative assessment of how close $\hat{B}_D$ and $\hat{B}_M$ are, albeit without the additional support of a p-value. For illustration the histograms for the three model fits corresponding to the largest value of $\hat{\omega}$ are shown in figure 6.1.

To allow a fuller appreciation of table 6.4, we present below the detailed results of a single model fit: the simple (misspecified) random effects model fit to the data set giving, under the true model, the median value of $\hat{\omega}$. For this model we obtained $\hat{\sigma^2} = 0.0892$ with an approximate standard error of 0.049. The data-based (D) and model-based (M) estimates of B as well as B_1, B_2, B_3 and B_4 are given below (table 6.2). The respective simulation-based standard errors are shown in brackets, in addition to the ones based on the exact formula. Note that for $\hat{B}_D$ we have given two simulation-based standard errors. The first was computed using formula (6.2.6). An alternative estimate of the same quantity is given on the line below, calculated as

$$\text{s.e.}(\hat{B}_D) = \sqrt{\sum_{i=1}^{50} \widehat{\text{Var}}(\hat{d}_{D_i})}, \tag{6.3.2}$$

	$\bar{B}$	$\bar{B}_1$	$\bar{B}_2$	$\bar{B}_3$	$\bar{B}_4$
D	7206.0	271.1	694.1	3261.1	2979.7
	$(4167.9)_S$	$(81.9)_S$	$(435.3)_S$	$(1509.3)_S$	$(3654.7)_S$
	$(3978.8)_S$				
	$(3700.4)_E$	$(76.0)_E$	$(467.9)_E$	$(1450.4)_E$	$(3371.1)_E$
M	10843.4	58.1	694.4	2746.1	7344.9

Table 6.2: Detailed results of simple random effects model fit for $\hat{\omega} = 0.907$; simulation-based and exact standard errors are subscripted by S and E respectively.

where $\widehat{\text{Var}}(\hat{d}_{D_i})$ is computed using (6.2.10). Both estimates are in reasonably good agreement with the one obtained from the exact variance formula, though it appears in this model fit and in others that the simulation-based standard error computed using (6.3.2) tends to be slightly closer to the true value; hence we use it for computing the simulation-based standardized statistic $\hat{S}_B$. At any rate, the agreement between 'exact' and 'simulated' standard errors seen in the table above is reflective of that observed in other model fits and other data sets, and is encouraging, given the modest number of simulations performed. From the values above we can directly compute $\hat{S}_B$, $\hat{S}_{B_1}$, $\hat{S}_{B_2}$, $\hat{S}_{B_3}$ and $\hat{S}_{B_4}$. These are given in table 6.3, and are flagged by (S) or (E), depending on which quantity was used to standardize.

	$\hat{S}_B$	$\hat{S}_{B_1}$	$\hat{S}_{B_2}$	$\hat{S}_{B_3}$	$\hat{S}_{B_4}$
(S)	-0.91	2.60	0.00	0.34	-1.19
(E)	-0.98	2.80	0.00	0.36	-1.29

Table 6.3: Standardized statistics computed from the values in table 6.2.

From table 6.4 we note that in all the data sets, the logistic model provides a much poorer fit than either of the random effects models. In the first two data sets ($\hat{\omega} = 0.001$ and $\hat{\omega} = 0.392$) there is no distinguishable difference in fit between the misspecified and the true

			$\tilde{S}_B$	$\tilde{S}_{B_1}$	$\tilde{S}_{B_2}$	$\tilde{S}_{B_3}$	$\tilde{S}_{B_4}$
$\hat{\omega} = 0.001$	Logistic Model	(S)	14.97	1.08	7.18	4.32	13.15
		(E)	15.11	0.97	6.90	4.07	13.66
	Misspecified R. E. Model	(S)	0.0030	-0.074	1.04	-0.77	0.18
		(E)	0.0033	-0.081	1.02	-0.79	0.20
	Correct R. E. Model	(S)	0.0049	-0.078	1.07	-0.77	0.18
		(E)	0.0054	-0.083	1.02	-0.79	0.20
$\hat{\omega} = 0.392$	Logistic Model	(S)	6.63	0.40	4.37	8.09	1.76
		(E)	6.34	0.34	4.23	7.67	1.68
	Misspecified R. E. Model	(S)	-0.32	-0.17	0.85	1.23	-1.03
		(E)	-0.34	-0.15	0.74	1.33	-1.12
	Correct R. E. Model	(S)	0.12	-0.96	0.42	1.56	-0.82
		(E)	0.12	-0.83	0.40	1.56	-0.85
$\hat{\omega} = 0.907$	Logistic Model	(S)	14.38	7.72	5.66	14.51	6.49
		(E)	14.12	7.71	5.39	13.62	6.49
	Misspecified R. E. Model	(S)	-0.91	2.60	0.00	0.34	-1.19
		(E)	-0.98	2.80	0.00	0.36	-1.29
	Correct R. E. Model	(S)	-0.11	0.40	-0.54	0.98	-0.62
		(E)	-0.11	0.40	-0.51	1.02	-0.64
$\hat{\omega} = 1.683$	Logistic Model	(S)	11.38	13.07	8.54	8.04	5.10
		(E)	11.05	11.91	8.48	7.92	4.91
	Misspecified R. E. Model	(S)	-1.16	4.70	1.13	-0.41	-1.33
		(E)	-1.24	4.92	1.02	-0.42	-1.44
	Correct R. E. Model	(S)	0.24	0.65	0.45	0.48	-0.25
		(E)	0.25	0.74	0.41	0.49	-0.27
$\hat{\omega} = 9.869$	Logistic Model	(S)	8.42	18.73	12.97	4.23	2.32
		(E)	8.22	15.85	12.51	3.97	2.30
	Misspecified R. E. Model	(S)	-2.04	5.23	0.94	-1.43	-1.84
		(E)	-2.24	5.15	0.85	-1.48	-2.04
	Correct R. E. Model	(S)	-0.66	-0.15	0.42	-0.56	-0.71
		(E)	-0.68	-0.16	0.39	-0.56	-0.75

Table 6.4: Analysis of $\tilde{S}_B = (\hat{B}_D - \hat{B}_M)/\sqrt{\widehat{\text{Var}}(\hat{B}_D)}$ based on simulated data from model (6.3.1); (each set of three horizontal panels corresponds to one data set – statistics standardized by simulation-based standard errors are flagged by (S), those computed on the basis of the exact formula by (E)).

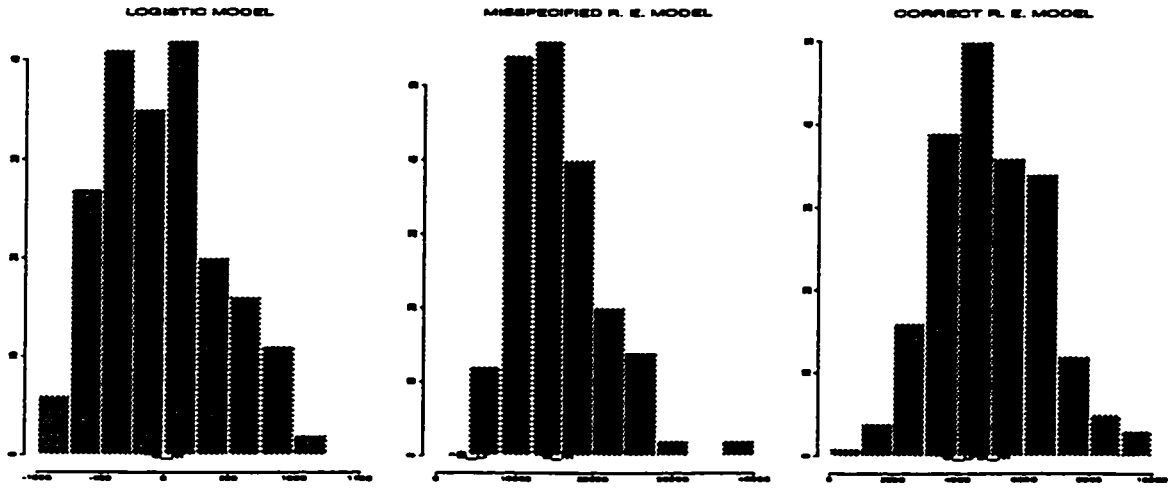
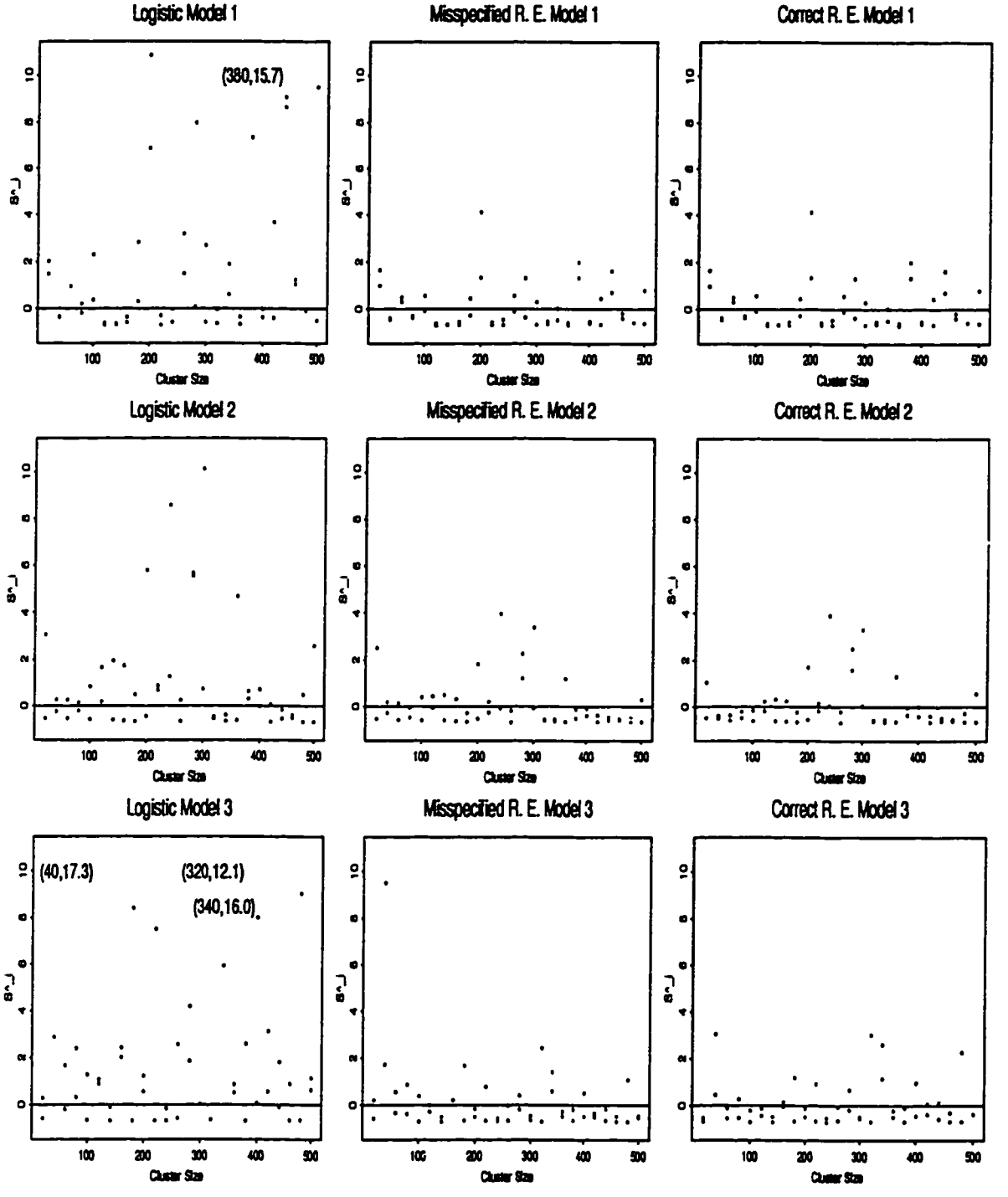


Figure 6.1: Histograms of the simulated values $\{\hat{B}_D^*\}$ for the three model fits corresponding to $\hat{\omega} = 9.869$; (for the logistic model, $\hat{B}_D = 3856.3$ is too far in the right tail to be located).

random effects model. Although there is strong evidence that the data are overdispersed, there is no basis for suggesting that this overdispersion is a function of cluster size. In the third data set ($\hat{\omega} = 0.907$) lack of fit in the misspecified model becomes apparent. Overall $|\hat{B}_D - \hat{B}_M|$ does not differ significantly from zero, but the model clearly seems to underestimate the dispersion in the smaller clusters ($n_i \leq 80$). Lack of fit of this type becomes more dramatically evident in the last two data sets ($\hat{\omega} = 1.683$ and $\hat{\omega} = 9.869$), with only the fit of the true model continuing to be adequate. Note that in the last three data sets the values $\hat{S}_{B_1}$ through $\hat{S}_{B_4}$ are decreasing monotonically for the misspecified model fit, with $\hat{S}_{B_1} > 0$ and $\hat{S}_{B_4} < 0$. This reflects the fact that, as anticipated, this model underestimates the dispersion in smaller clusters and overestimates it in larger ones.

For each of the 15 model fits under consideration, figure 6.2 shows plots of the standardized values $\hat{S}_i = (\hat{d}_{D_i} - \hat{d}_{M_i})/\sqrt{\widehat{\text{Var}}(\hat{d}_{D_i})}$ for each cluster, plotted against cluster size. These complement the numerical assessment given in table 6.4. As suggested above, the plots for the two random effects model fits to the first two data sets are very similar. The lack of fit of the misspecified model to the third data set seems due to one cluster in


Logistic Model 1
Misspecified R. E. Model 1
Correct R. E. Model 1
(380,15.7)
Logistic Model 2
Misspecified R. E. Model 2
Correct R. E. Model 2
Logistic Model 3
Misspecified R. E. Model 3
Correct R. E. Model 3
(40,17.3)
(320,12.1)
(340,16.0)
Cluster Size

(continued)

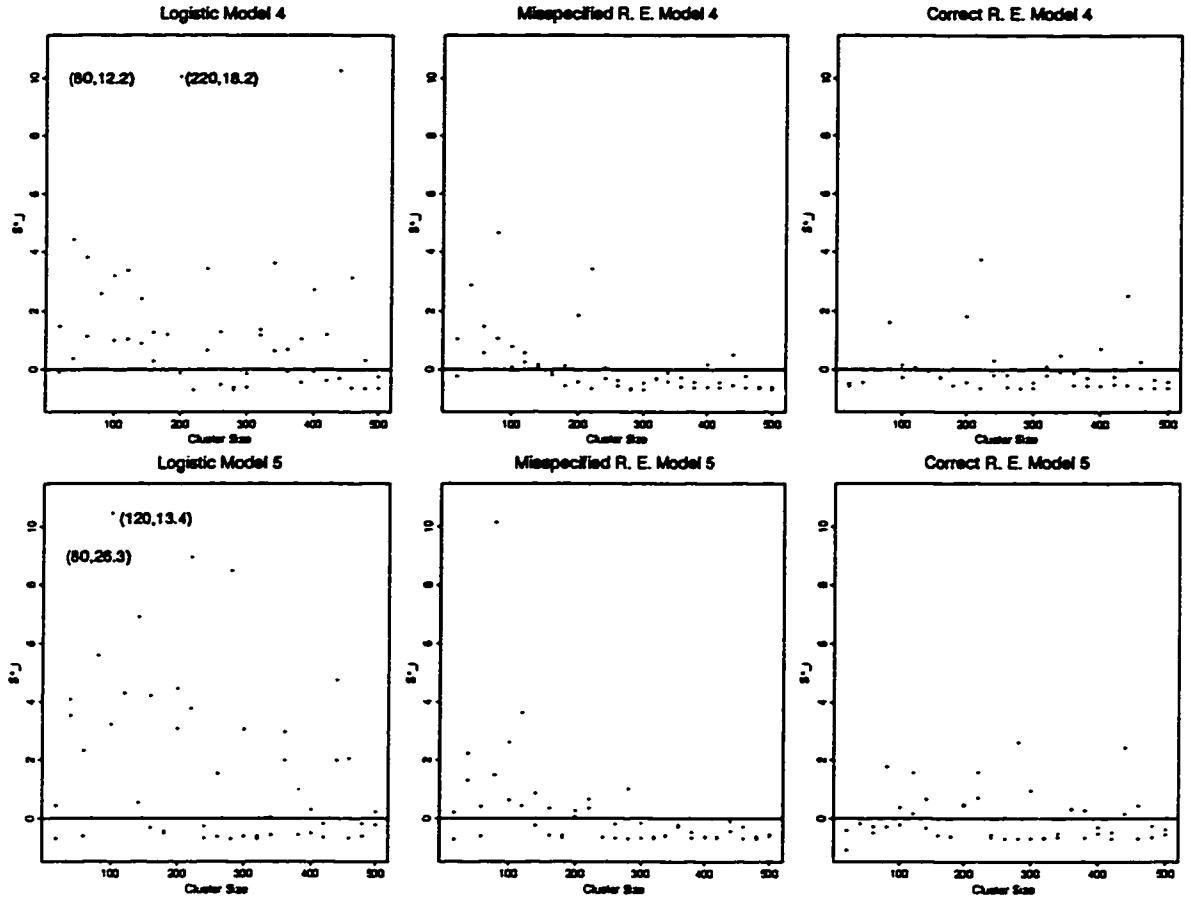


Figure 6.2: Plots of $\hat{S}_i$ vs Cluster Size, for simulated data sets 1 through 5 (corresponding to $\hat{\omega} = 0.001$ through 9.869) and various model fits; coordinates indicated refer to points out of range.

particular, whose variability, assuming the estimated mean to be accurate, is much larger than that predicted by the model. In the last two data sets the fit of the simple random effects model is such that $\hat{S}_i$ follows a noticeable decreasing pattern with increasing cluster size, indicating a likely dependence of the extraneous variation in the data on cluster size. In these as well as the third data set, the plots corresponding to the correct model fits, though still indicative of some lack of fit in individual clusters, show visible improvements over the misspecified random effects model fits.

	Logistic Model	L. R. Stat.	Misspecified R. E. Model	L. R. Stat.	Correct R. E. Model
$\hat{\omega} = 0.001$	-6251.96		-6217.02 ($\hat{\alpha} = -2.73$ (.30))		-6217.02 ($\hat{\alpha} = -2.72$ (1.71)) ($\hat{\gamma} = 0.0032$ (.63))
		69.88		0.0	
$\hat{\omega} = 0.392$	-6156.42		-6144.19 ($\hat{\alpha} = -3.35$ (.37))		-6143.01 ($\hat{\alpha} = -0.89$ (1.20)) ($\hat{\gamma} = 0.96$ (.48))
		24.46		2.36	
$\hat{\omega} = 0.907$	-6508.15		-6467.46 ($\hat{\alpha} = -2.42$ (.30))		-6463.73 ($\hat{\alpha} = -0.052$ (.84)) ($\hat{\gamma} = 0.97$ (.34))
		81.38		7.46	
$\hat{\omega} = 1.683$	-6097.97		-6067.32 ($\hat{\alpha} = -2.46$ (.32))		-6061.07 ($\hat{\alpha} = 0.54$ (0.79)) ($\hat{\gamma} = 1.31$ (.33))
		61.30		12.5	
$\hat{\omega} = 9.869$	-6228.48		-6191.29 ($\hat{\alpha} = -1.96$ (.29))		-6178.93 ($\hat{\alpha} = 2.29$ (0.91)) ($\hat{\gamma} = 1.98$ (.40))
		74.38		24.72	

Table 6.5: Values of the maximized log-likelihoods from the model fits presented in table 6.4. Point estimates of $\alpha = \log \sigma^2$ and γ, with standard errors in brackets, are shown where applicable. Likelihood ratio statistics for adjacent model comparisons are also given.

Since we were able to carry out maximum likelihood estimation for these models, and since the three model formulations we considered are all special cases of the true model (6.3.1), we briefly compare the results obtained here with likelihood ratio tests. Table 6.5 gives the values of the maximized log-likelihoods from the model fits presented in table 6.4. In all cases the maximized likelihood under the misspecified random effects model is much larger than that under the simple logistic model. No significant increase in likelihood is noted for the first two data sets, when comparing the fits of the misspecified and the correct random effects model. For the last three data sets, however, the maximized likelihood under the true model is indeed significantly larger when compared to the value under the

competing simpler random effects model; computing twice the difference in log-likelihood yields 7.46, 12.5 and 24.72 respectively for these data sets, suggesting a decidedly better fit under the true model.

In the following section we apply the methods developed here to a real data set, for which we compare two GEE models to decide on the best-fitting correlation structure.

6.4 Example: Analysis of Toxicology Data (Ganio and Schafer (1992))

Here we will investigate the toxicology data described in Ganio and Schafer (1992), and discussed briefly in section 6.2.1. Recall that 40 tanks of rainbow trout embryos were exposed to one of two compounds, aflatoxin B1 (AB1) or aflatoxicol (Al) at one of five doses, the response of interest being the incidence of liver tumors (r/n) seen in each of the tanks after one year. The data are listed below.

Dose (ppm)	Aflatoxin B1				Aflatoxicol			
0.010	3/86	5/86	4/88	2/86	9/87	5/86	2/89	9/85
0.025	14/87	14/90	9/83	12/88	30/86	41/86	27/86	34/88
0.050	29/90	31/89	33/89	26/87	54/89	53/86	64/90	55/88
0.100	44/86	40/80	44/89	43/88	71/88	73/89	65/88	72/90
0.250	62/87	67/88	59/88	58/84	66/86	75/82	72/81	73/89

In studies such as this one often finds extraneous tank-to-tank variation. In this particular set of data, however, there is evidence to suggest that there is some overdispersion in the aflatoxicol tanks only, and that the aflatoxin B1 tanks may in fact be underdispersed. We therefore examine three models for these data: the first, a logistic model fitting the main effects of compound type and dose, as well as their interaction; the second, a GEE model having the same marginal mean and assuming exchangeable correlation between fish

in the same tank, common across all tanks (GEE-M1). The third model also assumes that intra-cluster correlation is exchangeable, but estimates a separate correlation parameter for each group of tanks, allowing for differing dispersion under the two compound types (GEE-M2).

Letting ρ be the common correlation parameter in GEE-M1, and (ρ_{AB1}, ρ_{Al}) the correlation parameters for the AB1 and Al tanks respectively in GEE-M2, the following estimates were obtained:

$$\text{GEE-M1}: \quad \hat{\rho} = -0.00269; \qquad \text{GEE-M2}: \quad \hat{\rho}_{AB1} = -0.00821$$
$$\hat{\rho}_{Al} = 0.00284$$

To compute relevant simulation-based statistics as outlined in the previous sections, we need to be able to generate both positively and negatively correlated binomial data. The beta-binomial distribution can be used to generate positively correlated data. Simply use the estimated marginal probability of response and the positive intra-cluster correlation estimate for a given tank, say $\hat{p}_i$ and $\hat{\rho}_{Al}$, to calculate $a_i = \hat{p}_i(\frac{1}{\hat{\rho}_{Al}} - 1)$ and $b_i = (1-\hat{p}_i)(\frac{1}{\hat{\rho}_{Al}} - 1)$; then generate an observation p_i^* from the beta-binomial(n_i, a_i, b_i) distribution, where n_i refers to the tank size. Finally, use this value to simulate an observation R_i^* from the Binomial(n_i, p_i^*) distribution.

The generation of negatively correlated binomial data is more problematic. Emrich and Piedmonte (1991) describe an algorithm for generating vectors of binary variates which allows for a small range of negative correlations, but this range turns out to be too restrictive in this case, given our estimates of ρ and ρ_{AB1} and the relatively large cluster sizes. For similar reasons, the extended beta-binomial model presented in Prentice (1986) is not suitable either for simulating negatively correlated data in the present setting. We therefore resort to an *ad hoc* approach for simulating observations R_i which can be represented as

follows:

$$R_i = \sum_{j=1}^{n_i} Y_{ij}, \quad Y_{ij} \sim \text{Bin}(1, p_i), \quad \text{Corr}(Y_{ij}, Y_{ij'}) = \rho_i < 0, \quad j \neq j',$$

which implies that

$$\begin{aligned} \text{E}(R_i) = n_i p_i \quad \text{and} \quad \text{Var}(R_i) &= n_i p_i (1-p_i)[1 + \rho_i(n_i - 1)] \\ &< n_i p_i (1-p_i). \end{aligned}$$

Now given p_i and ρ_i it is easy to generate values

$$R_i' = S_i\sqrt{1 + \rho_i(n_i - 1)} + n_i p_i (1 - \sqrt{1 + \rho_i(n_i - 1)}),$$

where $S_i \sim \text{Bin}(n_i, p_i)$. Furthermore, R_i' has the same mean and variance as R_i. Unfortunately R_i' is not integer-valued, and hence cannot be directly used in place of R_i. However rounding to the nearest integer will produce a variable with approximately the correct specifications in terms of mean and correlation. Therefore, given the estimated marginal probability of response and the negative intra-cluster correlation estimate for a particular tank, say $\hat{p}_i$ and $\hat{\rho}$, generate a value S_i^* from the Binomial$(n_i, \hat{p}_i)$ distribution and use

$$R_i^* = S_i^*\sqrt{1 + \hat{\rho}(n_i - 1)} + n_i \hat{p}_i (1 - \sqrt{1 + \hat{\rho}(n_i - 1)})$$

rounded to the nearest integer as the simulated observation. The simulation-based standard errors computed using data generated in this manner were quite close to those based on formula (6.2.4); see table 6.6.

For each of the three models of interest we computed $\hat{B}_D$ and $\hat{B}_M$, as well as the breakdown of these terms into the sum of the contributions from the two types of tanks. Representing this division of B as $B = B_{AB1} + B_{Al}$, table 6.6 gives the data-based and

		$\hat{B}$	$\hat{B}_{AB1}$	$\hat{B}_{Al}$
Logistic Model	D	-177.00 $(137.14)_S$ $(140.42)_B$	-214.68 $(100.50)_S$ $(101.17)_B$	37.68 $(93.32)_S$ $(97.39)_B$
	M	0.00	0.00	0.00
GEE-M1	D	-177.05 $(104.19)_S$ $(107.99)_B$	-214.68 $(76.64)_S$ $(77.83)_B$	37.63 $(70.59)_S$ $(74.86)_B$
	M	-134.80	-67.74	-67.06
GEE-M2	D	-176.96 $(129.39)_S$ $(124.89)_B$	-214.67 $(30.41)_S$ $(30.09)_B$	37.71 $(125.77)_S$ $(121.21)_B$
	M	-135.78	-206.53	70.75

Table 6.6: Goodness-of-fit analysis for models fit to toxicology data given in Ganio and Schafer (1992); simulation-based standard errors for $\hat{B}_M$ were computed using formula (6.3.2) applied to these data.

model-based estimates of B, B_{AB1} and B_{Al} for each model. As before, standard errors are given in brackets. The simulation-based values were computed from 200 simulated data sets, as in section 6.3. Table 6.7 reports the resulting standardized statistics $\hat{S}_B$, $\hat{S}_{B_{AB1}}$ and $\hat{S}_{B_{Al}}$.

Fitting the logistic model to these data produces a deviance of 31.93 on 30 degrees of freedom, indicating at first glance a perfectly adequate fit. Indeed, even $\hat{B}_D$ and $\hat{B}_M$ do not differ significantly for this model, nor for either of the GEE models. However, decomposing $\hat{B}_D$ and $\hat{B}_M$ into the contributions from the 20 AB1 tanks and the 20 Al tanks, it is apparent that the logistic model does not adequately reflect the underdispersion in the data from the AB1 tanks. The data-based covariance contributions to $\hat{B}_D$ from this group tend to be significantly smaller than the covariance or fully model-based ones (these all being zero in the logistic model). GEE-M1 does not yield a much better fit, but GEE-M2 does show a

	$\bar{S}_B$	$\bar{S}_{B_{AB1}}$	$\bar{S}_{B_{Al}}$
Logistic Model	-1.26	-2.12	0.39
GEE-M1	-0.39	-1.89	1.40
GEE-M2	-0.33	-0.27	-0.27

Table 6.7: Standardized goodness-of-fit statistics, computed from the values in table 6.6; the 'exact' standard errors were used throughout to standardize.

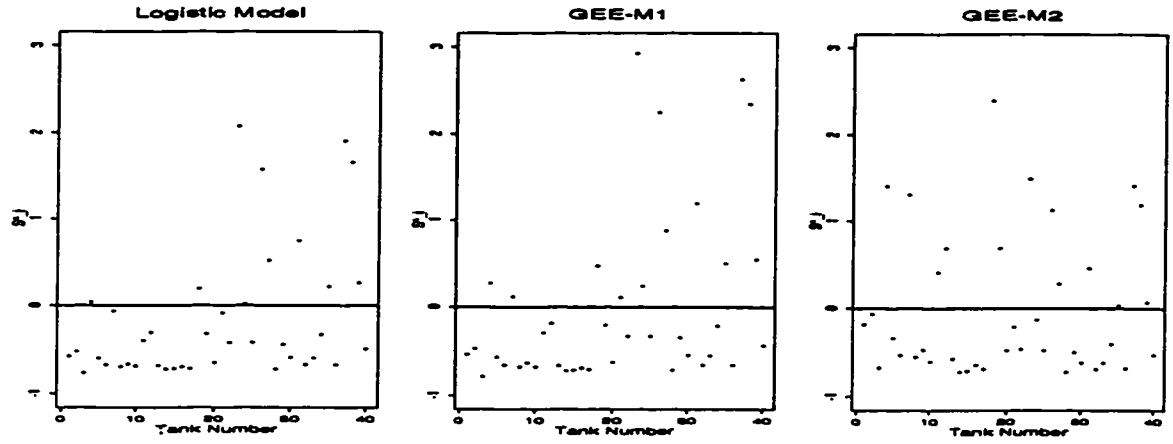


Figure 6.3: Plots of $\hat{S}_i$ vs Tank Number, for various model fits to Ganio and Schafer's toxicology data.

significant improvement in that under this model the model-based estimates are consistent with the data-based estimates, both across all tanks and when broken down by compound type.

Figure 6.3 shows the plots of the standardized values $\hat{S}_i = (\hat{d}_{D_i} - \hat{d}_{M_i})/\sqrt{\widehat{\text{Var}}(\hat{d}_{D_i})}$ for each tank, plotted against tank number where numbers 1 through 20 correspond to the AB1 tanks, and 21 through 40 correspond to the Al tanks. The lack of fit of the logistic model and GEE-M1 is manifested here in that for most of the AB1 tanks, $\hat{S}_i < 0$, implying that $\hat{d}_{D_i}$ is consistently smaller that $\hat{d}_{M_i}$; in other words the variance in this group of

tanks is consistently overestimated by both the the logistic model and GEE-M1. This pattern is no longer evident in the plot for GEE-M2, the more random scatter of points suggesting that the estimation of two correlation parameters more adequately reflects the correlation structure in the data. One should note that the plot for the logistic model fit looks qualitatively very similar to a plot of the deviance contributions from the individual tanks versus tank number. Such analogous deviance plots are not available, however, for the two GEE models.

6.5 Discussion

We have described a procedure to assess the fit of a model to correlated data under quite general conditions. Specifically, our focus was to test how well the covariance structure of a given model fits the empirical covariance structure as reflected by the data, assuming the model correctly specifies the mean.

Throughout this chapter it is essential to interpret the standardized test statistics as conditional on the estimates $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{\rho}}$ from the fit of the model under consideration. The statistic $\hat{S}_B$ (equation (6.2.7)) and the various analogous quantities discussed in section 6.2.3 are standardized assuming that $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{\rho}}$ are fixed values, i.e. taking into account only the variability in $\hat{B}_D$ (the data-based term). For a given model of interest, $\mathcal{M}$, $\hat{S}_B$ should therefore be used to address question

> *Q1: Are the observed data consistent with model $\mathcal{M}$, having true parameter values $\boldsymbol{p} = \hat{\boldsymbol{p}}$ and $\boldsymbol{\rho} = \hat{\boldsymbol{\rho}}$?*

as opposed to

> *Q2: Is the difference between $\hat{B}_D$ and $\hat{B}_M$ significantly different from zero, assuming the data were generated according to model $\mathcal{M}$ with parameters $\boldsymbol{p} = \hat{\boldsymbol{p}}$ and $\boldsymbol{\rho} = \hat{\boldsymbol{\rho}}$?*

In answering the first question, $\hat{B}_D$ is compared to a fixed constant whereas in the second question, the difference between two estimates of the same quantity is compared to zero. In addressing Q2 the variability in the difference $(\hat{B}_D - \hat{B}_M)$, which is a function of the variability in both $\hat{B}_D$ and $\hat{B}_M$, needs to be considered. $\text{Var}(\hat{B}_D - \hat{B}_M)$ can be estimated using a slightly altered version of the algorithm given on page (172) 171. Keeping steps (1), (2) and (4) the same, replace (3) by

3′. Fit model $\mathcal{M}$ to $\boldsymbol{Y}^*$ to obtain $\hat{\boldsymbol{p}}^*$ and $\hat{\boldsymbol{\rho}}^*$. Compute

$$\hat{\Delta}^* = B_D(\boldsymbol{Y}^*; \hat{\boldsymbol{p}}^*) - B_M(\hat{\boldsymbol{p}}^*; \hat{\boldsymbol{\rho}}^*).$$

This gives rise to $\hat{\Delta}_1^*, \hat{\Delta}_2^*, ..., \hat{\Delta}_N^*$ and hence the parametric bootstrap estimate

$$\widehat{\text{Var}}(\hat{B}_D - \hat{B}_M) = \frac{\sum_{s=1}^{N} \{\hat{\Delta}_s^* - \overline{\Delta}^*\}^2}{N-1}, \tag{6.5.1}$$

where $\overline{\Delta}^* = \sum_{s=1}^{N} \hat{\Delta}_s^*/N$. From (3′) it is clear that both $B_D(\boldsymbol{Y}^*; \hat{\boldsymbol{p}}^*)$ and $B_M(\hat{\boldsymbol{p}}^*; \hat{\boldsymbol{\rho}}^*)$ depend on $\boldsymbol{Y}^*$. Furthermore these two variables will be strongly positively correlated, hence we might expect that

$$\begin{aligned} \text{Var}(\hat{B}_D - \hat{B}_M) &= \text{Var}(B_D(\boldsymbol{Y}^*; \hat{\boldsymbol{p}}^*)) + \text{Var}(B_M(\hat{\boldsymbol{p}}^*; \hat{\boldsymbol{\rho}}^*)) - 2\text{Cov}(B_D(\boldsymbol{Y}^*; \hat{\boldsymbol{p}}^*), B_M(\hat{\boldsymbol{p}}^*; \hat{\boldsymbol{\rho}}^*)) \\ &< \text{Var}(B_D(\boldsymbol{Y}^*; \hat{\boldsymbol{p}}^*)). \end{aligned} \tag{6.5.2}$$

The inequality in (6.5.2) will hold true if ρ_B, the correlation between $B_D(\boldsymbol{Y}^*; \hat{\boldsymbol{p}}^*)$ and $B_M(\hat{\boldsymbol{p}}^*; \hat{\boldsymbol{\rho}}^*)$, exceeds 0.5. In fact

$$\rho_B > 0.5$$

is a conservative bound, since in order for (6.5.2) to hold we must have

$$\text{Var}(B_M(\hat{\boldsymbol{p}}^*; \hat{\boldsymbol{\rho}}^*)) - 2\rho_B \sqrt{\text{Var}(B_D(\boldsymbol{Y}^*; \hat{\boldsymbol{p}}^*))\text{Var}(B_M(\hat{\boldsymbol{p}}^*; \hat{\boldsymbol{\rho}}^*))} < 0,$$

implying that

$$\rho_B > \frac{1}{2}\sqrt{\frac{\text{Var}(B_M(\hat{\boldsymbol{p}}^*;\hat{\boldsymbol{\rho}}^*))}{\text{Var}(B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}}^*))}}.$$

But $\text{Var}(B_M(\hat{\boldsymbol{p}}^*;\hat{\boldsymbol{\rho}}^*)) \leq \text{Var}(B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}}^*))$, since

$$\begin{aligned}
\text{Var}(B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}}^*)) &= \text{Var}[\text{E}(B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}}^*)|\ (\hat{\boldsymbol{p}}^*,\hat{\boldsymbol{\rho}}^*))] + \text{E}[\text{Var}(B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}}^*)|\ (\hat{\boldsymbol{p}}^*,\hat{\boldsymbol{\rho}}^*))] \\
&= \text{Var}(B_M(\hat{\boldsymbol{p}}^*;\hat{\boldsymbol{\rho}}^*)) + \text{E}[\text{Var}(B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}}^*)|\ (\hat{\boldsymbol{p}}^*,\hat{\boldsymbol{\rho}}^*))] \\
&\geq \text{Var}(B_M(\hat{\boldsymbol{p}}^*;\hat{\boldsymbol{\rho}}^*)).
\end{aligned}$$

Hence $\sqrt{\text{Var}(B_M(\hat{\boldsymbol{p}}^*;\hat{\boldsymbol{\rho}}^*))/\text{Var}(B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}}^*))} \leq 1$ and $\rho_B > 0.5$. Therefore if $\widehat{\text{Var}}(\hat{B}_D)$ from (6.2.6) roughly equals the sampling variability in $B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}}^*)$, we could also expect that

$$\widehat{\text{Var}}(\hat{B}_D) > \widehat{\text{Var}}(\hat{B}_D - \hat{B}_M)$$

if $\rho_B > 0.5$. Hence, the test statistics developed in this chapter will be conservative in addressing the second question given above.

To illustrate, consider the goodness-of-fit summaries listed in table 6.4; we reanalyze the data set corresponding to $\hat{\omega} = 9.869$, using $\sqrt{\widehat{\text{Var}}(\hat{B}_D - \hat{B}_M)}$ as the standard error in computing $\hat{S}_B$, and computing $\hat{S}_{B_1}$, $\hat{S}_{B_2}$, $\hat{S}_{B_3}$ and $\hat{S}_{B_4}$ in a similar fashion, using a sample variance of the form (6.5.1) computed over the appropriate groups of clusters. Table 6.8 reports the simulation-based standard errors and those based on the exact formula (6.2.4), used to compute the test statistics for each of the three models for $\hat{\omega} = 9.869$ shown in table 6.4. These are flagged by (S) and (E) as before. Rows flagged by (A) report the alternative standard errors based on (6.5.1). The corresponding test statistics are given in table 6.9. As before computations were based on 200 simulated data sets.

For the logistic model, standardizing on the basis of (6.5.1) produces results very similar to those given in table 6.4. This is a reflection of the fact that $B_M(\hat{\boldsymbol{p}}^*;\hat{\boldsymbol{\rho}}^*) = 0$

Standard Errors Proposed for:			$\hat{S}_B$	$\hat{S}_{B_1}$	$\hat{S}_{B_2}$	$\hat{S}_{B_3}$	$\hat{S}_{B_4}$
$\hat{\omega} = 9.869$	Logistic Model	(S)	457.38	29.07	114.59	221.63	382.83
		(E)	469.38	34.35	118.78	236.21	386.30
		(A)	443.85	27.77	103.77	214.14	370.79
	Misspecified R. E. Model	(S)	5368.55	87.87	534.25	1959.99	4968.56
		(E)	4898.16	89.30	591.98	1894.63	4477.04
		(A)	3417.31	83.71	469.85	1654.51	3995.71
	Correct R. E. Model	(S)	1584.75	407.70	640.09	843.70	1106.19
		(E)	1544.33	378.05	674.68	843.00	1037.39
		(A)	505.31	298.48	494.59	645.61	740.29

Table 6.8: Comparison of standard errors proposed for computing $\hat{S}_B$, $\hat{S}_{B_1}$, $\hat{S}_{B_2}$, $\hat{S}_{B_3}$ and $\hat{S}_{B_4}$, for the data set corresponding to $\hat{\omega} = 9.869$ (see section 6.3); rows flagged by (S) and (E) denote simulated and exact standard errors as previously described, those flagged by (A) the alternative standard errors based on (6.5.1).

			$\hat{S}_B$	$\hat{S}_{B_1}$	$\hat{S}_{B_2}$	$\hat{S}_{B_3}$	$\hat{S}_{B_4}$
$\hat{\omega} = 9.869$	Logistic Model	(S)	8.42	18.73	12.97	4.23	2.32
		(E)	8.22	15.85	12.51	3.97	2.30
		(A)	8.69	19.61	14.32	4.38	2.39
	Misspecified R. E. Model	(S)	-2.04	5.23	0.94	-1.43	-1.84
		(E)	-2.24	5.15	0.85	-1.48	-2.04
		(A)	-3.21	5.49	1.07	-1.70	-2.28
	Correct R. E. Model	(S)	-0.66	-0.15	0.42	-0.56	-0.71
		(E)	-0.68	-0.16	0.39	-0.56	-0.75
		(A)	-2.07	-0.20	0.54	-0.73	-1.05

Table 6.9: Comparison of test statistics $\hat{S}_B$, $\hat{S}_{B_1}$, $\hat{S}_{B_2}$, $\hat{S}_{B_3}$ and $\hat{S}_{B_4}$, standardized according to the values in table 6.8.

under the assumption of independence, making (6.5.1) and (6.2.6) almost identical in this case. As anticipated however, the standard errors computed from (6.5.1) are smaller than those previously obtained for the two random effects models. This results in an inflation of the statistics $\hat{S}_B$, especially for the correct model. The more flexible a given model-based covariance structure, the better it is able to reflect the data-based covariance; hence ρ_B should be an increasing function of model complexity. The empirical correlation between $B_D(\boldsymbol{Y}^*;\hat{\boldsymbol{p}}^*)$ and $B_M(\hat{\boldsymbol{p}}^*;\hat{\boldsymbol{\rho}}^*)$ over the 200 simulated data sets was 0.782 under the misspecified random effects model and 0.944 under the correct model.

Note that the difference in the standard errors for each of $\hat{S}_{B_1}$, $\hat{S}_{B_2}$, $\hat{S}_{B_3}$ and $\hat{S}_{B_4}$ is not nearly as dramatic as for the global statistic $\hat{S}_B$. This is so because the contribution to the model-based estimate $B_M(\hat{\boldsymbol{p}}^*;\hat{\boldsymbol{\rho}}^*)$ from a specific group of clusters which are similar in their correlation structure, (say, $\hat{B}^*_{M_\ell}$), tends to be less variable across simulated data sets than $B_M(\hat{\boldsymbol{p}}^*;\hat{\boldsymbol{\rho}}^*)$ in its entirety. The difference in the standard errors for $\hat{S}_{B_\ell}$ computed on the basis of (6.5.1) and (6.2.6) diminishes as $\mathrm{Var}(\hat{B}^*_{M_\ell}) \longrightarrow 0$. Hence distinguishing between the two questions posed at the beginning of this discussion is less important when considering the fit of a modelled correlation stucture to the various parts of a division of the covariate space. Reasonable arguments can be made for supporting either question in the development of goodness-of-fit procedures. Proceeding as we have in this chapter has the advantage that simulation-based standard errors are more easily computed, since refitting the model of interest to each simulated data set is avoided. More importantly, it is desirable when possible to avoid simulation altogether, and as described in section 6.2.2 a closed-form analytic expression is available for (6.2.6). This is applicable when covariates are on the level of the cluster only, and also with individual-level covariates, provided cluster sizes are small. In contrast, no analogous exact variance expression exists for (6.5.1).

Assessing goodness-of-fit in the manner described in this chapter focusses attention on the covariance structure of the data in particular. The methods presented here complement those of Chapter 5 and are useful in general for examining the adequacy of a model for correlated data. Addressing the correlation structure in a basic fashion is central to making correct inferences, but refinements to the model for $\mathrm{Cov}(\boldsymbol{Y})$ can also be important. As indicated in Chapter 5, these may give useful insight into the mechanism which generated the data, and will impact on interval estimates for estimated probabilities. We have described procedures for examining this aspect of model fit; these should be interpreted as tests of the difference in a data-based and a model-based quantity, wherein the null hypothesis states that the true model is the one under consideration, with parameters equal to the estimates obtained from the fit of the model to the original data.

Chapter 7

Illustrations from the WSPP3 Data

In this chapter we focus on applications of the methods discussed in this thesis by considering various models for the WSPP3 data. The reader is referred back to Chapter 2 for a description of the study.

7.1 Examination of Grade 7 and 8 Smoking Behaviour in Baseline Non-Smokers

7.1.1 A Quasi Empirical Bayes Model

We begin by examining some aspects of elementary school smoking behaviour, as revealed by the WSPP3. Consider data from the first three years of the study, corresponding to the time the cohort of students spent in elementary schools. The subset of observations we selected includes the responses (self-reported smoking status) in grades 7 and 8 ($t = 1$ and 2, respectively) of those students who were non-smokers at baseline. Considering a complete-case analysis, this corresponded to 2 observations on each of 3380 students, attending a total of 99 schools. We examined a logistic model formulation expressing the

probability of smoking at time t as a function of the following variables, which were found to be of most relevance:

Cond : study condition (*Cond* = 1 for schools in one of the four treatment conditions and 0 otherwise);

Risk : an individual-level smoking risk score (*Risk* = 1 for students classified on the basis of external factors to be at low risk for smoking, *Risk* = 2 for students at medium risk and 3 for students at high risk);

Gr8surv : a school-level risk score, coded as a continuous covariate ranging between 0 and 100, with larger values indicative of higher-risk schools; this was derived from an examination of the proportion of smokers among the senior students in each school;

Gr8 : a grade effect (*Gr8* = 1 for a grade 8 observation, 0 otherwise).

In addition the interaction between *Cond* and *Gr8surv* (*C*×*Gr8surv*) was taken into consideration. Since students' individual-level risk could change over time, the value reported at time $t-1$ was used to predict the observation at time t. In this analysis we focussed on marginal smoking rates at each time point, relegating a student to the smoking state ($Y_{it}=1$) if (s)he reported to be either an experimental or a regular smoker, and to the non-smoking state ($Y_{it}=0$) otherwise. Letting x_{it} refer to the realization of the covariate vector (1, *Cond*, *Risk*, *Gr8surv*, *Gr8*, *C*×*Gr8surv*) for student i at time t, the results of fitting the composite QEB model

$$Y_{it}|\, b_{(it)} \sim \text{Bin}\left(1, \frac{\exp\{x'_{it}\beta + b_{(it)}\}}{1+\exp\{x'_{it}\beta + b_{(it)}\}}\right), \quad \text{Corr}(Y_{i1}, Y_{i2}|\, \{b_{(i1)}, b_{(i2)}\}) = \rho_{12},$$
$$b_{(it)} \in \{b_1, ..., b_{99}\}, \quad b_k \sim N(0, \sigma^2), \quad k = 1, ..., 99$$
$$i = 1, ..., 3380, \quad t = 1, 2$$

	Logistic		GEE		Emp. Bayes		QEB	
Term	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Int'cpt	-4.4741	(.2060)	-4.1339	(.2185)	-4.4237	(.2817)	-4.1424	(.2810)
Cond	0.4123	(.1855)	0.4352	(.2026)	0.3546	(.2803)	0.3773	(.2787)
Risk	0.8657	(.0536)	0.6734	(.0558)	0.8517	(.0547)	0.6896	(.0571)
Gr8surv	0.0284	(.00881)	0.0304	(.0096)	0.0287	(.0135)	0.0312	(.0134)
Gr8	0.8640	(.0806)	0.8560	(.0680)	0.8690	(0.812)	0.8641	(.0708)
C×*Gr8surv*	-0.0240	(.00992)	-0.0272	(.0108)	-0.0255	(.0149)	-0.0280	(.0147)
ρ_{12}			0.2857				0.2499	
σ^2					0.2001		0.1590	

Table 7.1: Various model fits to the WSPP3 elementary school data.

are given in table 7.1, along with the estimates from the three models which are special cases of the more general formulation, namely the ordinary logistic fit, assuming independence between all observations, the GEE fit, ignoring the random school effects, and the standard empirical Bayes logistic-normal model, assuming repeated observations on the same individual to be independent.

As anticipated, the standard errors for the two school-level covariates *Cond* and *Gr8surv* are similar for the empirical Bayes and composite model fits, and underestimated by the other models. The standard error for the individual-level covariate *Risk* is similar in the GEE and composite model fits; in this case it does not appear significantly understated by the other models, which is perhaps not surprising given the rather moderate estimate of ρ_{12} and only two observations per subject.

All models suggest that a student's individual risk score and that of the school(s) he/she attends are highly predictive of smoking status. Both the empirical Bayes and the composite model fits indicate that there is a marginally significant interaction between *Cond* and *Gr8surv*, suggesting that the intervention program may effect lower smoking rates in students, but only in the high-risk schools. A graphical representation of the results of fitting the QEB model, illustrating this phenomenon, is shown in figure 7.1. Both the

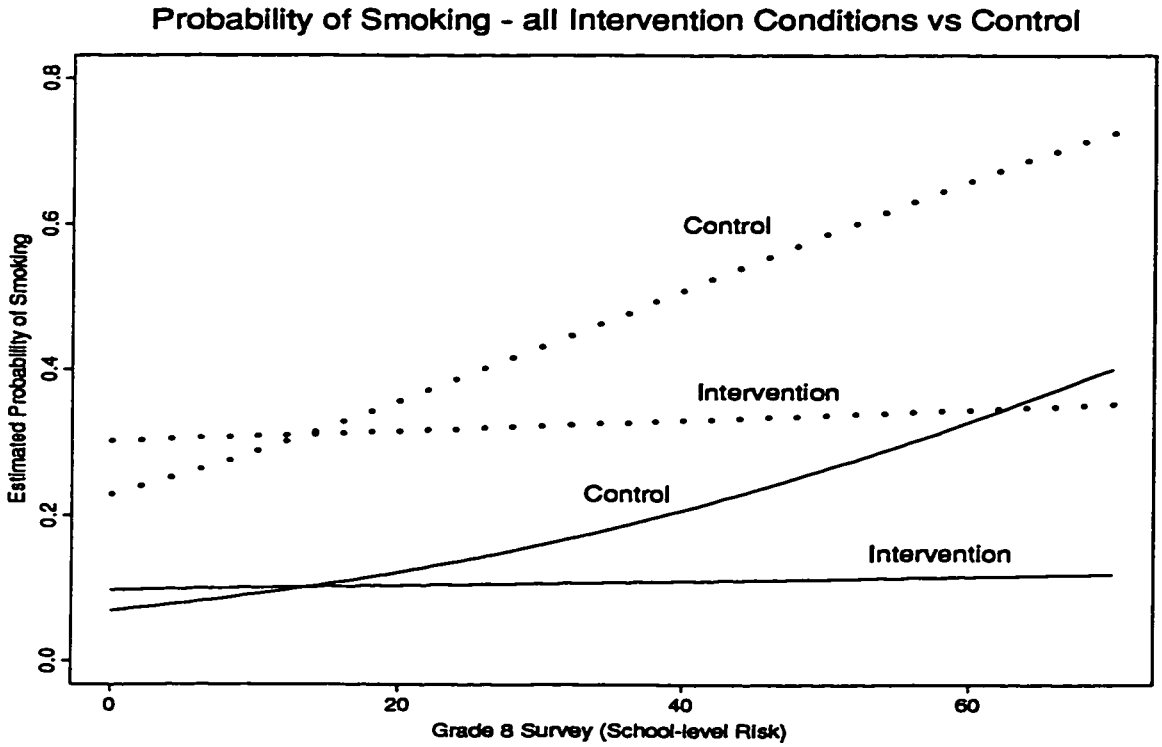


Figure 7.1: Estimated probability of smoking in grade 8, for high-risk individuals (*risk*=3) - dotted line, and low-risk individuals (*risk*=1) - solid line.

logistic and GEE model fits lead to a similar conclusion, but estimate this interaction to be more significant than it is in actual fact.

The standard errors reported for the QEB model are the model-based values. Robust standard errors were also computed, using the technique described in section 4.3. However, the last diagonal element of $\hat{I}_{QB}^{-1}$ (refer to equation (4.3.10)), corresponding to an estimate of $\text{Var}(\hat{\sigma^2})$, turned out to be negative. This problem had not occurred in any of the simulated data sets examined in Chapter 4; in trying to determine its cause we noted that the value of ν (equation (4.3.6)) computed from the QEB model was 62.6. This value is not unreasonable given the large number of clusters, but makes no sense in its interpretation as the mean of a hypothetical hyperprior distribution for σ^2, given that σ^2 is estimated to be about 0.15. However if we follow the same development as in section 4.3, but assume a

gamma hyperprior for $c\sigma^2$ instead of σ^2, for a positive constant c, then ν retains the same interpretation as the mean of the hyperprior and is given by

$$\nu = 2 + K + \hat{c\sigma^2} - \frac{\sum_{k=1}^{K} \hat{b}_k^2}{\hat{\sigma^2}}.$$

For c large enough, ν and $c\sigma^2$ are within a more reasonable range and the variance estimates from $\hat{I}_{QB}^{-1}$ are all positive. (For example for $c = 1000$, $\nu = 211.9$ and $\hat{c\sigma^2} = 149.4$). The robust variance estimates for the fixed effects parameters were found to be quite insensitive to the particular choice of c, and in fact these estimates were virtually identical to the model-based ones. This is not astonishing in light of the findings in Chapter 4 and the fact that σ^2 is estimated from quite a large number of clusters in this case.

7.1.2 A Closer Look at School-to-School Variability

We next examine more carefully the school-to-school variability seen in these data. For this purpose we will focus on random effects models as discussed in Chapter 5, ignoring the intra-individual correlation between grade 7 and grade 8 observations. We will however consider models including the same covariates as those in table 7.1.

We are interested first of all in determining whether the extraneous school-to-school variation can be ascribed to a particular group of students, on the basis of their individual-level smoking risk. To this end we begin by fitting a simple logistic model containing the predictors (*Cond*, *Risk*, *Gr8surv*, *Gr8*, *C*×*Gr8surv*), and constructing plots as described in section 5.4.2. In this case we consider three subgroups within each cluster, defined by the low, medium and high-risk observations. Letting

$$r_{kg} = \sum_{j|Y_{kj} \in \text{ group } g} Y_{kj}$$

be the total smoking response in school k out of n_{kg} observations for group g, $g = 1, 2, 3$,

we plot the residuals

$$\hat{\varepsilon}_{kg} = \log(\frac{r_{kg}}{n_{kg} - r_{kg}}) - \sum_{j|Y_{kj} \in \text{ group } g} x_{kj}'\hat{\beta}/n_{kg}, \quad k = 1, ..., 99, \quad g = 1, 2, 3$$

against the order statistics obtained by sorting the schools according to overall observed smoking proportion; see figure 7.2. We observe a wider spread on average in the residuals for students in the extreme risk groups (low or high) than for those who are at medium risk of smoking. When we combine low and medium risk students and compare this group to the high risk students (see the lower three plots in figure 7.2), we also see a wider spread in the residuals for the latter group as compared to the former.

In consideration of these plots we consider the general random effects model

$$\log\left(\frac{p_{kj}}{1 - p_{kj}}\right) = x_{kj}'\beta + \exp(\gamma_1 z_{kj1} + \gamma_2 z_{kj2}) \cdot b_k, \qquad b_k \sim N(0, \sigma^2) \tag{7.1.1}$$

for the response probability for the jth observation in school k, where x_{kj} is the corresponding covariate vector including an intercept and the predictors mentioned above, and

$$z_{kj1} = \begin{cases} 1 & \text{for a low-risk obs'n } (Risk = 1) \\ 0 & \text{otherwise,} \end{cases} \qquad z_{kj2} = \begin{cases} 1 & \text{for a high-risk obs'n } (Risk = 3) \\ 0 & \text{otherwise.} \end{cases}$$

We compare the fit of this model to the two special cases obtained by assuming $\gamma_1 = 0$ and $\gamma_1 = \gamma_2 = 0$. The results of these fits, along with that of the logistic model, are shown in table 7.2. In these random effects models, we integrated to obtain the marginal likelihood and then proceeded with maximum likelihood estimation. The value of the maximized log-likelihood for each model is also given in table 7.2.

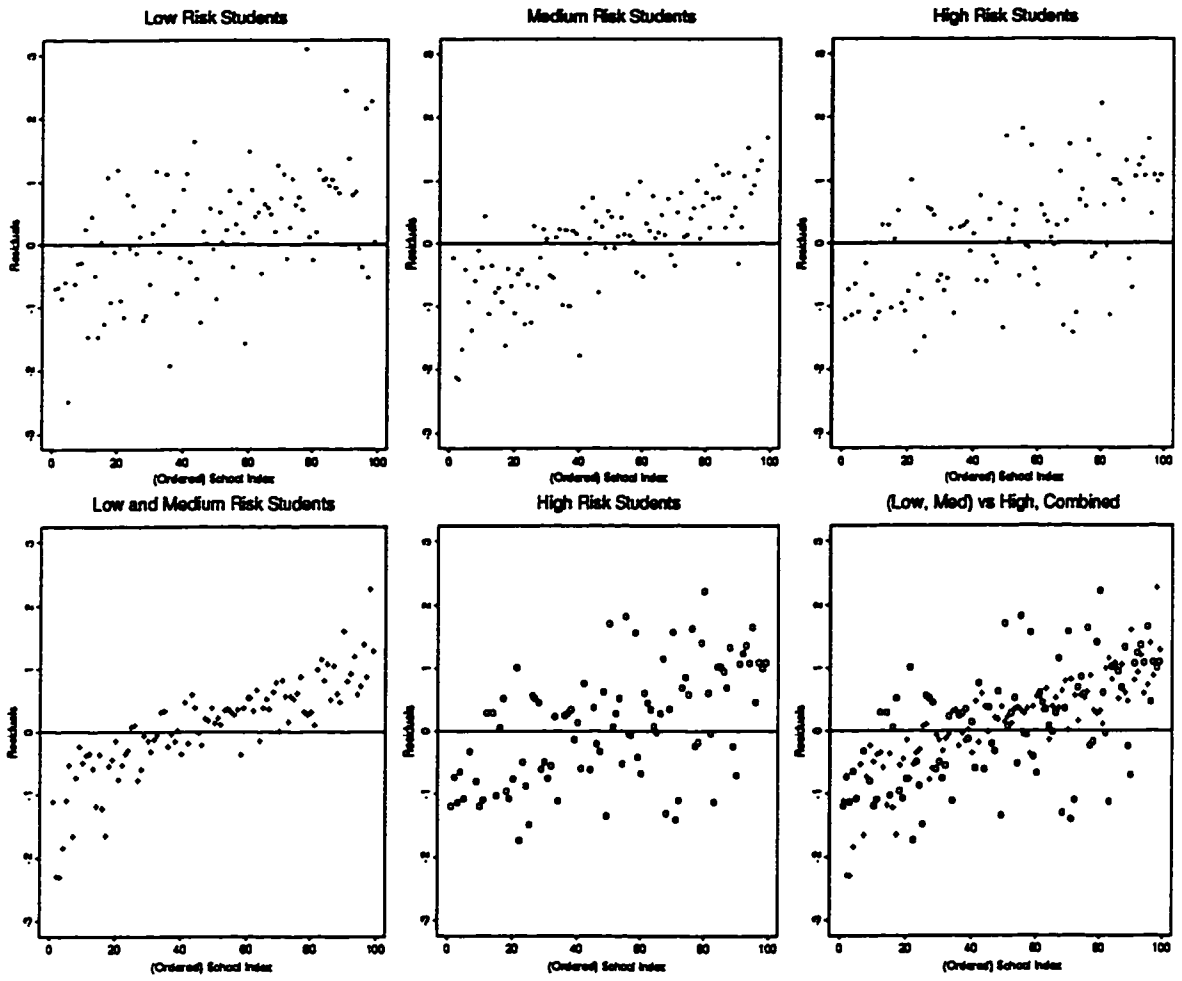


Figure 7.2: Exploratory residual plots for the WSPP3 elementary school data: investigating differences in overdispersion according to individual risk level.

In comparing the random effects model which estimates both γ_1 and γ_2 with the standard model assuming $f(z_{kj};\gamma)$ to be identically equal to 1, we note first of all that inferences for the covariates x_{kj} remain the same. Both parameter estimates and standard errors are similar in the two formulations. However the significant estimate of γ_2 in the more refined model suggests greater variability in the responses of the high-risk students as compared to those at medium and low risk. Combining the latter two risk groups (assuming γ_1 to be 0) and refitting the model yields the results in the last panel of table 7.2. We conclude

	Logistic		Random Effects Models					
Term	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Int'cpt	-4.4741	(.2060)	-4.4820	(.2820)	-4.3936	(.2760)	-4.3557	(.2720)
Cond	0.4123	(.1855)	0.3640	(.2802)	0.3135	(.2655)	0.3089	(.2668)
Risk	0.8657	(.0536)	0.8600	(.0550)	0.8398	(.0609)	0.8261	(.0597)
Gr8surv	0.0284	(.00881)	0.0293	(.0135)	0.0283	(.0128)	0.0278	(.0129)
Gr8	0.8640	(.0806)	0.8764	(.0816)	0.8711	(.0819)	0.8728	(.0819)
C×Gr8surv	-0.0240	(.00992)	-0.0261	(.0149)	-0.0248	(.0141)	-0.0251	(.0142)
γ_1					0.3184	(.3560)		
γ_2					0.7508	(.3159)	0.6335	(.2615)
σ^2			0.1950		0.0993		0.1300	
llik	-2321.25		-2302.76		-2299.93		-2300.31	

Table 7.2: Models for school-to-school variability in the WSPP3 elementary school data, based on formulation (7.1.1).

from this fit that the standard deviation of the random effects distribution governing the school-to-school variability in smoking rates among high-risk students is $e^{0.6335}$ or about twice as large as that for the remaining students. Adjusting for this difference in dispersion also produces a significant increase in log-likelihood over the standard random effects model.

As a final note, a word of caution is in order when interpreting residual plots such as those in figure 7.2. Care is needed especially when dividing clusters into groups of observations on the basis of a covariate which also has a significant impact on the response probability, as is the case in this application. Note that when the observed proportion of successes in a subgroup of a given cluster is quite small or even zero, a misleadingly large residual on the scale of the linear predictor can still be obtained even if the estimated proportion in this subgroup is reasonably close to the observed. As an extreme example of this note that $|\text{logit}(p_o) - \text{logit}(\hat{p})| \to \infty$ as $p_o \to 0$ for any $\hat{p} > 0$; in practice we may replace the observed value $\text{logit}(r/(n-r))$ with $\text{logit}((r+0.5)/(n-r+0.5))$ to avoid infinite values, but the stated problem persists. Thus it looks from the plots in figure 7.2 as though the

variability in the residuals for the low-risk students is just as large as that for the high-risk students; but since the probabilities in the former group are quite small compared to those in the latter group, the discrepancy between observed and fitted proportions is not nearly as dramatic. As a consequence, fitting a model comparing only low-risk students to all the rest (assuming γ_2 to be 0) yields a small and non-significant point estimate for γ_1.

Another question of interest is whether or not the school-to-school variability seen in these data is a function of school size. Knowing this would help one to distinguish between two competing theories postulated to explain the cause of the overdispersion. The first of these suggests that as a result of some common environmental effect acting on all students in a given school, there is a common (exchangeable) correlation between any two individuals in a given school. This intra-school correlation may be constant or may vary across schools, but it need not be a function of school size. In contrast, the second theory relates to peer cohesion and attributes extraneous school-to-school variability to the strong behavioral similarity among students in small peer groups within schools. The smaller the school, the fewer such subgroups it would contain and consequently the larger the relative variability in the responses. Similarly, this variability would be smaller in larger schools, containing many subgroups. In this case intra-school correlation would necessarily have to be a decreasing function of school size.

To investigate the dependence of school-to-school variability on school size in the WSPP3 data, consider fitting the random effects model

$$\log\left(\frac{p_{kj}}{1-p_{kj}}\right) = x'_{kj}\beta + \exp(\gamma n_k)\cdot b_k, \qquad b_k \sim N(0,\sigma^2), \tag{7.1.2}$$

in which the vector x_{kj} contains the same covariates as listed previously. A negative estimate of γ would imply a decrease in variability between schools as school size n_k increases and in a marginal sense would suggest that intra-school correlation decreases with increasing school size. The estimates obtained from the fit of this model and the model

	R. E. Model ($\gamma = 0$)		R. E. Model ($\gamma = \hat{\gamma}$)	
Term	est.	s.e.	est.	s.e.
Int'cpt	-4.4820	(.2820)	-4.5140	(.2701)
Cond	0.3640	(.2802)	0.4520	(.2727)
Risk	0.8600	(.0550)	0.8581	(.0551)
Gr8surv	0.0293	(.0135)	0.0274	(.0131)
Gr8	0.8764	(.0816)	0.8785	(.0817)
C×*Gr8surv*	-0.0261	(.0149)	-0.0272	(.0146)
γ			-0.00392	(.00251)
σ^2	0.1950		0.3903	
llik	-2302.76		-2301.40	

Table 7.3: Models for school-to-school variability in the WSPP3 elementary school data, based on formulation (7.1.2).

assuming $\gamma = 0$ (also given in table 7.2) are listed in table 7.3. Judging by the standard error of $\hat{\gamma}$ and the increase in likelihood when estimating γ as opposed to assuming it to be 0, there is mild evidence against the hypothesis of no association between school-to-school variability and school size. As a more direct way of assessing the extent to which the fit of (7.1.2) improves on that of the standard random effects model, we apply the methods of checking goodness-of-fit discussed in Chapter 6. Table 7.4 reports the estimates $\hat{B}_D$ and $\hat{B}_M$, as well as the standardized statistics $\hat{S}_B$ for the two models of interest. On the whole both models seem to adequately reflect the correlation structure; although the discrepancy between $\hat{B}_D$ and $\hat{B}_M$ is larger under the simpler random effects model than under the more general formulation, this difference is not significant. One does however see some improvement in the fit of the latter model when considering plots of the standardized values $\hat{S}_k$ plotted against school size; see figure 7.3. It is apparent that the simpler model tends to overestimate the dispersion among large schools and underestimate it among smaller ones. Further, this problem seems to be resolved in the fit of the second model

		$\hat{B}$	$\hat{S}_B$
R. E. Model ($\gamma = 0$)	D	930.53 (694.25)	-0.86
	M	1525.97	
R. E. Model ($\gamma = \hat{\gamma}$)	D	894.18 (346.82)	-0.27
	M	987.43	

Table 7.4: Goodness-of-fit analysis for models fit to WSPP3 elementary school data; simulation-based standard errors, based on 200 simulated data sets, are shown in brackets.

which estimates γ. Hence there appears to be at least some support for arguing that the variability between schools is a decreasing function of school size.

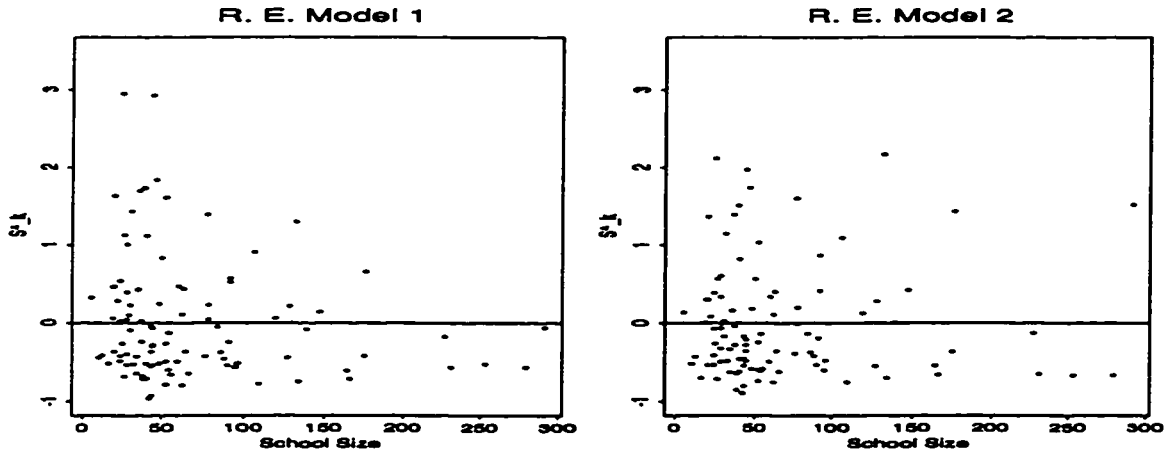


Figure 7.3: Plots of $\hat{S}_k$ vs School Size (n_k), for the models in table 7.3; Model 1: $\gamma = 0$, Model 2: $\gamma = \hat{\gamma}$.

Unfortunately this conclusion does not allow us to rule out either of the two behavioral theories discussed above; the absence of an inverse relationship between school-to-school variability and school size would have allowed dismissing the second of these, based on peer cohesion. It is worth pursuing somewhat the implications of this theory. Suppose one

models the correlation structure of the responses from a given school with an exchangeable correlation parameter, say ρ_k, when in actual fact only the responses among individuals in the same peer group are correlated, and students from different groups in the same school act independently. Assume for the sake of simplicity an exchangeable correlation ρ'_k between any two individuals in the same peer group, common for all groups in the school. It is instructive to examine the relationship between ρ_k and ρ'_k, which will depend on the school size and number of subgroups contained within the school. If covariates are on the level of the cluster only this is quite straightforward. Consider a model which assumes

$$R_k = \sum_{j=1}^{n_k} Y_{kj}, \qquad Y_{kj} \sim \text{Bin}(1, p_k), \qquad \text{Corr}(Y_{kj}, Y_{kj'}) = \rho_k, \quad j \neq j',$$

under which

$$\text{Var}(R_k) = n_k p_k(1-p_k) + n_k(n_k-1)\rho_k p_k(1-p_k) = n_k p_k(1-p_k)(1+\rho_k(n_k-1)).$$

Suppose however that $n_k = m_1 + \cdots + m_{s_k}$ and that the true structure of the data is as follows:

$$R_k = \sum_{r=1}^{s_k}\sum_{j=1}^{m_r} Y_{krj}, \qquad \text{Corr}(Y_{krj}, Y_{krj'}) = \rho'_k, \quad j \neq j', \; r = 1, ..., s_k,$$
$$\text{Corr}(Y_{krj}, Y_{kr'j'}) = 0, \quad r \neq r',$$

where Y_{krj} is the jth observation in the rth subgroup in school k. From this follows that

$$\text{Var}(R_k) = n_k p_k(1-p_k) + \sum_{r=1}^{s_k} m_r(m_r-1)\rho'_k p_k(1-p_k)$$

and hence the relation between ρ_k and ρ'_k which would equate these two expressions for

$m = 3$							
	s_k:	1	2	5	10	50	100
	(n_k:	3	6	15	30	150	300)
ρ'_k=	0.2	0.2000	0.0800	0.0286	0.0138	0.0027	0.0013
	0.4	0.4000	0.1600	0.0571	0.0276	0.0054	0.0027
	0.6	0.6000	0.2400	0.0857	0.0414	0.0081	0.0040
	0.8	0.8000	0.3200	0.1143	0.0552	0.0107	0.0054
$m = 5$							
	s_k:	1	2	5	10	50	100
	(n_k:	5	10	25	50	250	500)
ρ'_k=	0.2	0.2000	0.0889	0.0333	0.0163	0.0032	0.0016
	0.4	0.4000	0.1778	0.0667	0.0327	0.0064	0.0032
	0.6	0.6000	0.2667	0.1000	0.0490	0.0096	0.0048
	0.8	0.8000	0.3556	0.1333	0.0653	0.0129	0.0064

Table 7.5: Values of ρ_k computed from equation (7.1.3).

$\mathrm{Var}(R_k)$ is

$$\rho'_k = \frac{\rho_k n_k(n_k - 1)}{\sum_{r=1}^{s_k} m_r(m_r - 1)} \quad \Longleftrightarrow \quad \rho_k = \frac{\rho'_k \sum_{r=1}^{s_k} m_r(m_r - 1)}{n_k(n_k - 1)}. \tag{7.1.3}$$

Note that the there is no explicit dependence on p_k in the above equations, although if estimation proceeds by way of a cluster-specific model, $\hat{\rho}_k$ will in fact depend on $\hat{p}_k$. Table 7.5 reports the values of ρ_k which would correspond to each of several fixed values of ρ'_k and various choices of s_k and m, where we let $m = m_1 = \cdots = m_{s_k}$. One way to interpret a given estimate of intra-school correlation, therefore, would be to think of it as resulting from the presence of numerous small groups in the school, in each of which the smoking behaviour of all members is essentially the same.

To get some sense of how this might relate to the WSPP3 data, consider the fit of model (7.1.2). Although we also included individual-level covariates in this fit, we will just examine marginal correlation estimates using the cluster-level random effects model

(5.3.7) with $f(z_k;\hat{\gamma}) = \exp(-0.00392\ n_k)$ and $\hat{\sigma^2} = 0.3903$; n_k is in the range (5, 291), implying that $f(z_k;\hat{\gamma})$ varies from 0.320 to 0.981. We give several estimates of intra-school correlation in the table below, corresponding to various values of n_k and $p_k(0)$. (Recall that in (5.3.7) $p_k(b_k) = [1 + \exp(-(x_k'\beta + f_k \cdot b_k))]^{-1}$).

	n_k	10	20	50	100	200	300
$p_k(0)$ =	0.05	0.0214	0.0195	0.0148	0.0095	0.0041	0.0018
	0.10	0.0367	0.0337	0.0261	0.0172	0.0076	0.0034
	0.50	0.0771	0.0720	0.0585	0.0410	0.0196	0.0091

Throughout this development we have made strong and largely unverifiable simplifying assumptions. If the behavioural theory based on peer cohesion were indeed the sociological phenomenon causing the cross-sectional clustering in the WSPP3 data, it would likely apply under much more general conditions. The peer group sizes surely would not all be the same, nor would the specific intra-group correlation necessarily be constant across groups. It would for instance make sense given the previous results that peer groups consisting of purely high-risk students should be more strongly correlated in their smoking behaviour than groups with varied backgrounds. This is also intuitively clear.

7.2 Transition Models

In this section we consider transition models of the type discussed in section 3.5.4. There the notation Y_{ijt} was used to denote the response of individual j in cluster i at time point t. More generally, we can adopt the notation introduced in section 4.2.2 and used in 7.1.1 above, letting Y_{it} refer to the tth observation on individual i, and noting implicitly in which of the 99 schools this observation was taken.

We will consider two models: the first will consider the responses collected in grades 8 and 7 and condition each observation on the previous year's responses, the second will

consider the grade 8 data only and condition on the responses from the previous two years for each student. Such conditioning could largely remove the extraneous school-to-school variability; nevertheless we use random effects to adjust for any remaining overdispersion. In addition to the covariates defined in section 7.1, let *Y1* and *Y2* denote the response (smoking status) obtained from the cohort in grades 6 and 7 respectively. Letting $t = 3, 2, 1$ correspond to grades 8, 7 and 6, we specify logistic formulations for the probability that student i will be smoking at time t ($Y_{it} = 1$), either of the form

$$P(Y_{it} = 1 |\ y_{it-1}, b_{(it)}) = \frac{\exp(h(x_{it}, y_{it-1}) + b_{(it)})}{1 + \exp(h(x_{it}, y_{it-1}) + b_{(it)})}, \qquad t = 2, 3 \tag{7.2.1}$$

or

$$P(Y_{i3} = 1 |\ y_{i2}, y_{i1}, b_{(i3)}) = \frac{\exp(h(x_{i3}, y_{i2}, y_{i1}) + b_{(i3)})}{1 + \exp(h(x_{i3}, y_{i2}, y_{i1}) + b_{(i3)})}. \tag{7.2.2}$$

Here $h(\cdot)$ is used generically to denote a linear combination of covariates and past reponses, and possibly their interactions. Once again $b_{(it)}$ denotes a random effect associated with the school attended by individual i at time t, assumed to be normally distributed with mean 0 and some variance σ^2. Maximizing the integrated likelihood, the above models were fit as standard logistic-normal random effects models, taking the conditional responses to be independent of one another within a given school. Model summaries are given in table 7.6.

Conditioning on only the previous year's response, we note that smoking status at time $t - 1$ is highly predictive of the response at time t. The effect of the other predictors on the response is the same regardless of previous smoking status, with the exception that the impact of risk appears to be less dramatic for previous smokers as compared to non-smokers.

The results of the second model fit, conditioning the grade 8 observations on the grade 7 and 6 responses, are perhaps even more interesting. For students who were non-smokers

Conditioning Y_{it} on Y_{it-1}			Conditioning Y_{it} on Y_{it-1}, Y_{it-2}		
Term	est.	s.e.	Term	est.	s.e.
Int'cpt	-4.1613	(.2637)	*Int'cpt*	-3.5044	(.3234)
Cond	0.3389	(.2560)	*Cond*	0.5611	(.3244)
Risk	0.7193	(.0591)	*Risk*	0.6529	(.0721)
Gr8surv	0.0291	(.0122)	*Gr8surv*	0.0380	(.0153)
Gr8	0.6878	(.0784)	*C*×*Gr8surv*	-0.0481	(.0173)
C×*Gr8surv*	-0.0267	(.0135)	*Y2*	2.1534	(.1475)
Y_{t-1}	3.1412	(.3706)	*Y1*	7.1314	(1.802)
Y_{t-1}×*Risk*	-0.4544	(.1539)	*Y1*×*C*	-3.9646	(1.689)
			Y1×*Risk*	-0.9487	(.3099)
			Y1×*Gr8surv*	-0.1332	(.0590)
			Y1×*C*×*Gr8surv*	0.1558	(.0647)
σ^2	0.1404			0.2191	
llik	-2400.63			-1391.30	

Table 7.6: Transition models based on (7.2.1) and (7.2.2).

in grades 6 and 7, the interaction between *Cond* and *Gr8surv* already discussed in section 7.1.1 is quite pronounced, both in terms of the magnitude of the regression coefficient and its significance. The odds of a student smoking in grade 8 are estimated to be 8.6 times higher if that student was also smoking in grade 7. We also witness a very strong impact on the grade 8 responses, due to grade 6 smoking status. Not only is there a tremendous marginal effect if a student was also smoking in grade 6, but the impact of each of the covariates *Cond*, *Risk*, *Gr8surv* differs too. The interactions described by the parameter estimates in table 7.6 are interpreted in table 7.7 in terms of estimated probabilities of grade 8 smoking, given various covariate combinations. All entries in this table were computed assuming a medium level of individual risk, i.e. *Risk* = 2. We see that students in the intervention condition who were non-smokers in grade 6 fare better in high-risk schools

		Cond=1			*Cond*=0		
		Gr8surv=0	*Gr8surv*=60		*Gr8surv*=0	*Gr8surv*=60	
Y2=1							
	Y1=1	0.856	0.927	(53)	0.994	0.372	(14)
	Y1=0	0.626	0.477	(215)	0.489	0.903	(45)
Y2=0							
	Y1=1	0.409	0.594	(48)	0.954	0.064	(19)
	Y1=0	0.163	0.096	(2510)	0.100	0.520	(610)

Table 7.7: Some estimated probabilities of smoking, computed from the second model in table 7.6, conditioning grade 8 observations on grade 7 and 6 responses; all values are calculated assuming *Risk* = 2; sample sizes for each group of observations are given in brackets.

than students in similar control schools. However the opposite seems to be true for grade 6 smokers: the smoking rates seem to be lower in low-risk intervention as compared to control schools, but higher in the high-risk intervention schools. Of course one should keep in mind that there were only a small number of smokers in grade 6 when interpreting these findings. The sample sizes for each of the eight groups of observations distinguished in table 7.7 are given in brackets beside the estimated probabilities.

It is also surprising that conditioning on the previous response(s) does not seem to reduce the school-to-school variability; note for instance the similar estimates of σ^2 reported in tables 7.6 and 7.1. This may have in part to do with the variability in the number of first-time smokers in grades 7 and 8. Also, random effects in a logistic-normal model are additive on the logistic scale, and effects of similar size can have very differing impacts on estimated probabilities, depending on the fixed effects formulation for the mean.

To finish this section we examine the goodness-of-fit of the second model in table 7.6, compared to that of the nested logistic model, assuming $\sigma^2 = 0$. Table 7.8 reports the estimates $\hat{B}_D$ and $\hat{B}_M$ and the standardized statistics $\hat{S}_B$ for both the logistic and the random effects model. The logistic model is clearly inadequate in terms of describing the

		$\hat{B}$	$\hat{S}_B$
Logistic Model	D	496.03 (75.67)	6.56
	M	0.00	
R. E. Model	D	535.84 (313.88)	-0.50
	M	691.86	

Table 7.8: Goodness-of-fit for random effects transition model vs. logistic model.

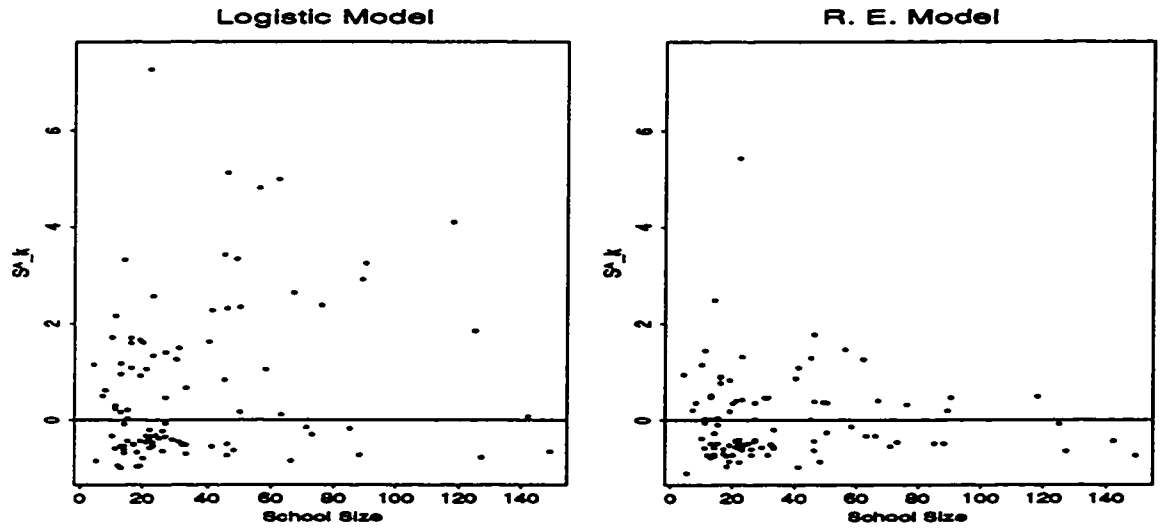


Figure 7.4: Plots of $\hat{S}_k$ vs School Size for the second random effects transition model in table 7.6 and the corresponding logistic model.

covariance structure of the data, whereas the random effects model does quite well. As in figure 7.3 we have plotted the standardized values $\hat{S}_k$ against school size in figure 7.4. In the plot for the random effects model only one school stands out from the rest as not well modelled. Closer examination revealed that out of the 22 grade 8 observations in this school, 8 (36 %) were smoking. This is not an unduly large proportion; however, none of these 22 students had been smokers in the previous two grades. The estimated mean probability of smoking for this school (0.12) therefore underestimates the observed proportion of smokers. Since the proposed goodness-of-fit procedures assume to begin with that the mean of the model is correctly specified, a much larger variability in the responses from this school would need to be postulated in order to reconcile the large discrepancy in observed and fitted values in this case.

7.3 Secondary School Smoking Behaviour

Here we consider data from the secondary phase of WSPP3 only. In order to assess the post-intervention impact of the elementary smoking prevention program, and any additional effect due to the highschool intervention, we examined the high school smoking behaviour of those students who were in one of the five original study conditions in grade 6, reported to be non-smokers in grade 8, attended one of the 30 study highschools in grade 9 and provided complete data until grade 12. This resulted in 4 observations (grades 9 through 12) on each of 1381 students, attending at any given time either one of the 30 study schools, or a non-study highschool; (a student could for instance transfer to a non-study school after grade 9). As in section 7.1.1 we examined logistic model formulations for the probability of student i smoking at time t. The covariates we retained in this case were the following:

Gr10, *Gr11*, *Gr12* : indicator variables taking value 1 for observations in grades 10, 11 and 12 respectively, and 0 otherwise,

Cond9 : highschool study condition (*Cond9* = 1 for intervention schools and 0 otherwise),

Cond6 : elementary school study condition (defined as *Cond* in section 7.1.1),

Gender : taking value 1 for female students, 0 for males,

Risk : individual-level smoking risk score (as defined previously).

In addition let $G\times C9$ denote the interaction between *Gender* and *Cond9*. Here we carried out an intent-to-treat analysis, treating individuals as though they remained in the same study condition throughout their highschool career. That is, the value of *Cond9* assigned to schools in grade 9, and hence to all students therein, was taken to be fixed for each student, even if the individual moved to a school of the opposite study condition or to a non-study school at a later time. We considered a comparison of the same four models as are listed in table 7.1, both for regressions including and not including indicators for time. In this case we are modelling Y_{it}, $i = 1,...,1381$, $t = 1,...,4$ ($\Leftrightarrow$ Gr.9,...,Gr.12), and hence estimate, for the GEE and QEB model fits, the 6 parameters in the intra-individual correlation matrix

$$\text{Corr}(Y_i) = \text{Corr}\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}.$$

In addition, the empirical Bayes and QEB fits model Y_{it} conditional on $b_{(it)}$, where in this case $b_{(it)} \in \{b_1,...,b_{31}\}$, $b_k \sim N(0,\sigma^2)$, $k = 1,...,31$. Thirty of these random effects correspond to the 30 study schools, and the 31st is specified for those observations taken in any other school; (since all students in this data set attended one of the 30 study schools in grade 9, such observations were necessarily responses in grade 10 or higher). Model summaries are provided in table 7.9.

	Logistic		GEE		Emp. Bayes		QEB	
Term	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Int'cpt	-2.1700	(.1123)	-1.8499	(.1479)	-2.2692	(.1474)	-1.7854	(.1700)
Cond9	-0.1078	(.0866)	-0.1311	(.1257)	-0.0987	(.1380)	-0.0781	(.1604)
Cond6	-0.0005	(.0769)	0.0281	(.1120)	0.0748	(.0931)	0.0620	(.1251)
Gender	-0.0734	(.0883)	-0.1225	(.1287)	-0.0702	(.0901)	-0.1014	(.1270)
Risk	0.8345	(.0391)	0.5412	(.0441)	0.8294	(.0399)	0.5044	(.0425)
G×C9	0.2408	(.1196)	0.3119	(.1738)	0.2191	(.1216)	0.2662	(.1715)
ρ_{12}			0.355				0.304	
ρ_{13} ρ_{23}			0.248	0.573			0.211	0.582
ρ_{14} ρ_{24} ρ_{34}			0.191	0.416 0.582			0.159	0.418 0.614
σ^2					0.1213		0.0794	
Int'cpt	-2.7800	(.1289)	-2.0922	(.1529)	-2.9068	(.1630)	-2.2699	(.1803)
Gr10	0.7640	(.0927)	0.7778	(.0691)	0.7789	(.0933)	0.7852	(.0726)
Gr11	0.9682	(.0916)	1.0407	(.0729)	0.9916	(.0925)	1.0474	(.0766)
Gr12	1.0515	(.0914)	1.1442	(.0754)	1.0796	(.0927)	1.1499	(.0790)
Cond9	-0.1112	(.0879)	-0.1236	(.1271)	-0.0946	(.1409)	-0.0812	(.1637)
Cond6	0.0002	(.0780)	0.0162	(.1128)	0.0760	(.0947)	0.0616	(.1287)
Gender	-0.0748	(.0898)	-0.0872	(.1299)	-0.0676	(.0916)	-0.0814	(.1311)
Risk	0.7680	(.0398)	0.3718	(.0423)	0.7644	(.0407)	0.4007	(.0442)
G×C9	0.2550	(.1214)	0.2747	(.1757)	0.2282	(.1236)	0.2563	(.1770)
ρ_{12}			0.410				0.375	
ρ_{13} ρ_{23}			0.332	0.585			0.298	0.546
ρ_{14} ρ_{24} ρ_{34}			0.283	0.426 0.580			0.253	0.390 0.535
σ^2					0.1323		0.0817	

Table 7.9: Various model fits to the WSPP3 secondary school data.

Examining briefly the standard errors of the regression coefficients, we note the similar values for individual-level covariates under the QEB and GEE fits, which are underestimated by the other two models (see in particular *Cond6* and *Gender*). *Cond9* as defined is also an individual-level covariate, though insofar as most students do remain in the same study condition over time, it should also behave like a school-level covariate. This is indeed the case. In both sets of models in table 7.9, the standard error of the coefficient for *Cond9* is inflated under the GEE and empirical Bayes fit as compared to the logistic fit, and the QEB model in each case provides the largest estimate of all four models.

No problems as discussed in section 7.1.1 were encountered when computing robust standard errors for these models; again these values were very similar to the model-based quantities, agreeing up to the third decimal place.

There is no discernible difference in the highschool smoking rates between students who had received the WSPP3 elementary intervention and those who had not. The secondary intervention also shows little impact. A previous analysis of the data suggested that males who were non-smokers in grade 8 and subsequently entered a secondary intervention school showed significantly lower smoking rates than females at the end of grade 10, and that this difference was maintained to the end of grade 12 by those males from high-risk elementary schools (Brown and Cameron (1997)). Nevertheless considering all the data over the entire span of the highschool observation period, the effects of intervention, gender and their interaction are slight. It would be worthwhile to consider separate analyses of specific portions of the data, to avoid unduly large models containing complicated higher-order interactions. For example one might examine males and females separately, and within each gender look at groups of students with similar risk profiles.

From the various model fits in table 7.9 we note that responses from the same individual over time tend to be more strongly correlated in later years; compare also to the estimates of the correlation between grade 7 and 8 observations from table 7.1. In addition, considering the models which include the grade indicators adjusting for time, we note that the odds of a

student smoking in grade 10 as compared to grade 9 are about $e^{0.79} = 2.2$ times larger, but that the analogous increase in comparing grades 11 and 10 is only a factor of $e^{1.05-0.79} = 1.3$, and almost negligible in comparing grades 12 and 11 (taken from the QEB model fit in table 7.9). This suggests that smoking behaviour in adolescents becomes more firmly set throughout the highschool years, i.e. less easily influenced by intervention programming. Launching such programs in much earlier grades seems to provide some measure of success, though it is not clear how to maintain these positive results as students move to secondary schools, apart from providing continued intensive intervention throughout highschool.

Chapter 8

Conclusion

The analysis of correlated binary data is a vast field. We have discussed a number of established modelling approaches for such data in this thesis, and have suggested contributions to facilitate the analysis of data such as that arising from the WSPP3. We addressed the problem of analyzing cluster-correlated longitudinal observations by combining the methods of empirical Bayes estimation for random effects models with generalized estimating equations, to obtain a single model formulation. We expect that data exhibiting such a composite correlation structure arise in various other settings as well; one might for instance consider an application to the analysis of multi-center longitudinal clinical trials, in which the different centers would form the cross-sectional clusters.

An interesting point to note from the simulations in Chapter 4 is the fact that when data are clustered cross-sectionally and also longitudinally correlated, and the effect of interest can be modelled in terms of a cluster-level covariate, then the intra-individual correlation need not necessarily be taken into account in order to obtain a valid estimate of this effect and its standard error. This has particularly important implications when designing a study such as WSPP3, in which a great administrative burden would be avoided if one were not required to collect personal information from each student in order to be able to

link repeated obervations over time to the same individual. At the same time one must be clear about the limitations involved in restricting oneself to a series of cross-sectional studies. In section 2.2 we outlined three general questions of interest surrounding a smoking prevention program, dealing with the effectivenenss of the intervention in general, the nature of smoking onset, and how specifically the intervention relates to smoking onset. The first of these questions can be addressed through school-level covariates only, and hence a cross-sectional investigation would suffice. On the other hand, studying smoking onset in students requires information on the smoking behaviour of individuals over time, necessitating a longitudinal study design. Individual-level characteristics, as well as school-level covariates, also play an important role in distinguishing differing treatment effects in subgroups of the cohort. In this case intra-individual correlation should be taken into account as well.

In Chapter 5 we explored at some length the relationship between population-averaged and cluster-specific models, and the usefulness of relatively straightforward random effects models to describe certain correlation stuctures. The models we discussed allow for the inclusion of covariate information in the specification of the random term of the linear predictor, which hinges on a deterministic function whose parameter(s) are estimated from the data. This approach allows the modelling of general features of the correlation structure in a parsimonious fashion, without having to appeal to more complicated models involving several variance components. The material in this chapter is complemented by that in Chapter 6, where we considered testing the goodness-of-fit of the covariance structure of a model, given that the mean is correctly specified. The methods proposed are quite general and hinge on comparing data-based and model-based covariance estimates in a similar manner as one might compare observed data and fitted values to assess the adequacy of a mean-model specification.

We ended by describing the results of several model fits to the WSPP3 data, using the techniques developed. In this smoking prevention program the highschool intervention

seems to have had little impact, but there is some evidence that the elementary school intervention was effective for students in high-risk schools. Interestingly there appeared to be no significant differences among the four treatment conditions described in Chapter 2; both nurses and teachers seemed to achieve about the same results, regardless also of training method. Furthermore, the content and style scores also mentioned in Chapter 2 were found to be inconsequential as well. Such information is useful in that it has important bearing on the design and implementation of future programs and studies.

The WSPP3 data motivated much of the work in this thesis, and was also used for the purpose of illustrations. However the thesis does not by any means constitute an exhaustive analysis of the data. The interested reader is referred to the final report by Brown and Cameron (1997) and in general to the Health Behaviour Research Group at the University of Waterloo.

References

Ashby, M., Neuhaus, M., Hauck, W., Bacchetti, P., Heilbron, D., Jewell, N., Segall, M., and Fusaro, R. (1992). An annotated bibliography of methods for analyzing correlated categorical data. *Statistics in Medicine* **11**, 67-99.

Bahadur, R. R. (1961). "A representation of the joint distribution of responses to n dichotomous items". In *Studies in Item Analysis and Prediction*, H. Solomon (ed.). Stanford Mathematical Studies in the Social Sciences VI, Stanford, California: Stanford University Press, 158-168.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The New S Language: a Programming Environment for Data Analysis and Graphics.* Pacific Grove, CA: Wadsworth and Brooks/Cole.

Best, J. A., Brown, K. S., Cameron, R., Manske, S. M., and Santi, S. (1995). Gender and predisposing attributes as predictors of smoking onset: implications for theory and practice. *Journal of Health Education* **26**, S52-S60.

Bowman, D. and George, E. O. (1995). A saturated model for analyzing exchangeable binary data: applications to clinical and developmental toxicity studies. *Journal of the American Statistical Association* **90**, 871-879.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear

mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Brown, K. S. and Cameron, R. (1997). Long-term evaluation of an elementary and secondary school smoking intervention. *Final Report to the NHRDP.*

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *American Statistician* **46**, 167-174.

Chan, J. S. K. and Kuk, A. Y. C. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* **53**, 86-97.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. New York: Wiley.

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

Conaway, M. (1990). A random effects model for binary data. *Biometrics* **46**, 317-328.

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

Cook, R. J. and Ng, E. T. M. (1997). A logistic-bivariate normal model for overdispersed two-state Markov processes. *Biometrics* **53**, 358-364.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical, B* **74**, 187-220.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika* **70**, 269-274.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

Crowder, M. J. (1978). Beta-binomial anova for proportions. *Applied Statistics* **27**, 34-37.

Crowder, M. J. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measurements. *Biometrika* **82**, 407-410.

Dean, C. B. (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistcal Association* **87**, 451-457.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the E M algorithm. *Journal of the Royal Statistical Society, B* **39**, 1-38.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Clarendon Press.

Dobson, A. J. (1990). *An Introduction to Generalized Linear Models.* London: Chapman and Hall.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**, 1-26.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**, 54-77.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician* **45**, 302-304.

Farrell, P. J. (1991). Empirical Bayes estimation of small area proportions. Ph.D. Dissertation, Department of Management Science, McGill University, Montreal.

Farrell, P. J., MacGibbon, B. and Tomberlin, T. J. (1994). Protection against outliers in empirical Bayes estimation. *Canadian Journal of Statistics* **22**, 365-376.

Farrington, C. P. (1996). On assessing goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statitical Society B* **58**, 349-360.

Fattinger, K. E., Sheiner, L. B. and Verotta, D. (1995). A new method to explore the distribution of interindividual random effects in non-linear mixed effects models. *Biometrics* **51**, 1236-1251.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistcal Association* **74**, 269-277.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51**, 309-317.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141-151.

Fitzmaurice, G. M. and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* **90**, 845-852.

Follman, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151-168.

Ganio, L. M. and Schafer, D. W. (1992). Diagnostics for overdispersion. *Journal of the American Statistical Association* **87**, 795-804.

George, E. O. and Bowman, D. (1995). A full likelihood procedure for analysing exchangeable binary data. *Biometrics* **51**, 512-523.

Gibbons, R. and Hedeker, D. (1994). Application of random-effects probit regression models. *Journal of Consulting and Clinical Psychology* **62**, 285-296.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics* **31**, 1208-1211.

Hamerle, A. (1990). On a simple test for neglected heterogeneity in panel studies. *Biometrics* **46**, 193-199.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley-Interscience.

Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics, Continuous Univariate Distributions*, Vol. 2. Boston: Houghton-Mifflin.

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

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association* **82**, 739-750.

Laird, N. M. and Ware, J. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963-974.

le Cessie, S. and van Houwelingen, H. C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics* **51**, 600-614.

Lefkopoulou, M., Moore, D., and Ryan, L. (1989). The analysis of multiple correlated binary outcomes: application to rodent teratology experiments. *Journal of the American Statistical Association* **84**, 810-815.

Liang, K.-Y. (1987). A locally most powerful test for homogeneity with many strata. *Biometrika* **74**, 259-264.

Liang, K.-Y. and Waclawiw, M. A. (1990). Extension of the Stein Estimating Procedure through the use of estimating functions. *Journal of the American Statistical Association* **85**, 435-440.

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

Liang, K-Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, B* **54**, 3-40.

Lindsey, J. K., and Lambert, P. (1997). On the appropriateness of marginal models for repeated measurements in clinical trials. *to appear.*

Lipsitz, S. R., Fitzmaurice, G. M. and Molenberghs, G. (1996). Goodness-of-fit tests for ordinal response regression models. *Applied Statistics* **45**, 175-190.

Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J., and Laird, N. M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics* **50**, 270-278.

Lipsitz, S. R., Fitzmaurice, G. M., Sleeper, L. and Zhao, L. P. (1995). Estimation methods for the joint distribution of repeated binary observations. *Biometrics* **51**, 562-570.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1990). Using the jackknife to estimate the variance of regression estimators from repeated measures studies. *Communications in Statistics - Theory and Methods* **19**, 821-45.

Longford, N. T. (1994). Logistic regression with random coefficients. *Computational Statistics and Data Analysis* **17**, 1-15.

MacGibbon, B. and Tomberlin, T. J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology* (Statistics Canada) **15**, 237-252.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics* **11**, 59-67.

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

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162-170.

Miller, M. E. (1995). Analysing categorical responses obtained from large clusters. *Applied Statistics* **44**, 173-186.

Molenberghs, G. and Ritter, L. L. (1996). Methods for analyzing multivariate binary data, with association between outcomes of interest. *Biometrics* **52**, 1121-1133.

Moore, D. F. (1987). Modelling the extraneous variance in the presence of extra-binomial variation. *Applied Statistics* **36**, 8-14.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78**, 47-55.

Murray, D., Rooney, B., Hannan, P., Peterson, A., Ary, D., Biglan, A., Botvin, G., Evans, R., Flay, B., Futterman, R., Getz, J., Marek, P., Orlandi, M., Pentz, M., Perry, C. and Schinke, S. (1994). Intraclass correlation among common measures of adolescent smoking: estimates, correlates, and applications in smoking prevention studies. *American Journal of Epidemiology* **140**, 1038-1050.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A* **135**, 370-384.

Neuhaus, J. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* **1**, 249-273.

Neuhaus, J., Hauck, W. and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755-762.

Neuhaus, J., Kalbfleisch, J. D. and Hauck, W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**, 25-35.

Neuhaus, J. and Segal, M. R. (1997). An assessment of approximate maximum likelihood estimators in generalized linear mixed models. Peer-reviewed article in Gregoire, T. et al., *Modelling longitudinal and spatially correlated data*. New York: Springer-Verlag, 11-22.

Ng, E. T. M. (1997). Statistical Inference for Heterogeneous Event History Data. Ph.D. Dissertation, Department of Statistics and Actuarial Science, University of Waterloo.

O'Hara-Hines, R. J. and Lawless, J. F. (1993). Modelling overdispersion in toxicological mortality data grouped over time. *Biometrics* **49**, 107-121.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics* **B23**, 939-951.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* **81**, 321-327.

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

Rai, S. N. and Matthews, D. E. (1993). Improving the E M Algorithm. *Biometrics* **49**, 587-591.

Rao, J. N. K. and Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics* **48**, 577-585.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association* **83**, 231-241.

Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485-497.

Royall, R. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review* **54**, 221-226.

Sitter, R. R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association* **87**, 755-765.

Sitter, R. R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics* **20**, 135-154.

Smith, P. J. and Heitjan, D. F. (1993). Testing and adjusting for departures from normal dispersion in generalized linear models. *Applied Statistics* **42**, 31-41.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Association B* **51**, 47-60.

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

Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika* **67**, 250-251.

Tukey, J. W. (1958). Bias and confidence in not quite large samples (abstract). *Annals of Mathematical Statistics* **29**, 614.

Vonesh, E. F., Chinchilli, V. M. and Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics* **52**, 572-587.

Waclawiw, M. and Liang, K.-Y. (1993). Prediction of random effects in the generalized linear model. *Journal of the American Statistical Association* **88**, 171-178.

Waclawiw, M. and Liang, K.-Y. (1994). Empirical Bayes estimation and inference for the random effects model with binary response. *Statistics in Medicine* **13**, 541-551.

Walker, S. (1996). A EM algorithm for non-linear random effects models. *Biometrics* **52**, 934-944.

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

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**, 949-952.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics* **31**, 144-148.

Williams, D. A. (1988). Extra-binomial variation in toxicology. Presented at the I.B.C. meetings in Namur, Belgium.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233-243.

Zeger, S. L. and Karim, M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79-86.

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

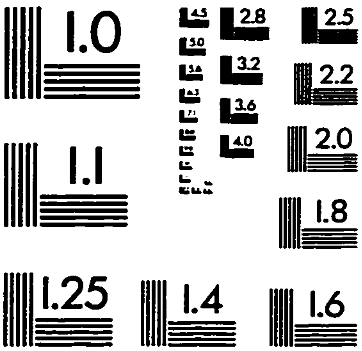
Zeger, S. L., Liang, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine* **11**, 1825-1839.

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

Zeger, S. L., Liang, K.-Y., and Self, S. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika* **72**, 31-38.

Zhao, L. P. and Prentice, R. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642-648.

IMAGE EVALUATION TEST TARGET (QA-3)



150mm

6″

APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989