Likelihood inference for exchangeable binary data with varying cluster sizes

Catalina Stefanescu

and

Bruce Turnbull*

School of ORIE, Cornell University, Ithaca, NY 14853, USA

SUMMARY

Various parameterizations of models for correlated exchangeable binary variables are discussed. We suppose that we have available a sample of $K(\geq 1)$ independent sets or “clusters” of such binary data. We propose use of the EM algorithm to obtain maximum likelihood estimates, especially when the clusters have varying sizes. The methodology is illustrated with applications to a study of familial disease aggregation and to the design of a proposed group randomized cancer prevention trial.

1. Introduction

This article was motivated by discussions concerning the design of the proposed *Prevention of Cancer with Selenium in Europe and America* (PRE-

* *Corresponding author's email address:* bruce@orie.cornell.edu

Key words: Clinical trials; EM algorithm; Exchangeable binary data; Trend test

CISE) clinical trial. The purpose of this randomized, double-blind, placebo-controlled trial with subjects from the United States and five European countries was to test the hypothesis that a daily nutritional supplement of selenium (Se) will reduce cancer incidence. The goal was to compare the health experience of the Se supplementation group with a placebo group with the primary endpoint being total cancer incidence within 5 years of randomization. The desire to have a high power of 90% to detect a small but important treatment effect of 10% in the 5-year cancer incidence, led to an initial proposal of a sample size of 52,000 subjects — Natarajan et al. (1996). The costs implied by such a large sample size alarmed some of the investigators. Thus, as a means to lower the costs of the study, a clustered sampling design was adopted, in which each study subject was asked to invite one or more other members of the same household to also participate, provided that the eligibility requirements were met. This would decrease the cost of the study in three ways. First, it would reduce the recruitment effort. Second, by the mutual reinforcement of the household members, it would help with compliance; thus any treatment effect would be less likely to be diluted by non-compliance. Third, it was thought that the same reinforcement phenomenon would lead to a lower dropout rate, also helping in the overall efficiency of the trial design.

One possible approach for treatment assignment was a paired or stratified design whereby treatment assignments were blocked by household and household members assigned different treatments. However, the use of this design was deemed impractical. First, logistically, the distribution of pills (placebo or Se), would be much more convenient if all members of the same house-

hold were assigned the same treatment. Second and more important, because many of the subjects would be elderly, it was felt that they might confuse the pills delivered to a household and thus receive the wrong treatment unless all in a household received the same treatment. Therefore instead, a cluster randomized design was proposed with households, varying in size between one and four persons, as the randomization units. Due to the similarity of lifestyle, nutritional, genetic and other possible factors, subjects sampled from the same household were considered to form a cluster. The sample size calculations should thus be amended to account for correlation between treatment effect outcomes within clusters.

Samples of correlated binary data arise commonly in many other application areas, such as studies of disease occurrence in families or repeated measurements on study subjects. We consider a subclass of multivariate binary distributions in which the responses may be assumed to be exchangeable. A sequence of binary random variables $Y_1, Y_2, \dots$ is exchangeable if

$$\Pr(Y_1 = y_1, \dots, Y_r = y_r) = \Pr(Y_1 = y_{\pi(1)}, \dots, Y_r = y_{\pi(r)})$$

for any r , any $(y_1, \dots, y_r)$ and any permutation π of $1, 2, \dots, r$. The assumption of exchangeability is reasonable in many family studies, developmental toxicity experiments, group randomized clinical trials, and studies of diseases of paired organs, for example. In these cases, the standard methods of analysis that ignore the intracluster correlation may lead to inefficient estimates and erroneous standard errors.

Previous inference methods for clustered exchangeable binary data include non-likelihood based techniques such as generalized estimating equations (e.g. Prentice (1988), Liang, Qaqish and Zeger (1992)) and pseudo-

likelihood (e.g. Connolly and Liang (1988)), as well as likelihood methods (e.g. Bowman and George (1995), Brooks et al. (1997)). Most of the previous approaches involve the mean response and the second order correlation. However, models with higher order interactions are often of interest. In particular, a saturated model in which arbitrary interactions of all orders are allowed, can be used to test adequacy of fit of more parsimonious nested submodels.

In the present paper we focus on maximum likelihood estimation for such models. When the data consist of clusters of equal sizes, estimation is quite straightforward. However, when clusters are of unequal sizes it is important to specify model parameters in such a way that they have a consistent interpretation whatever the cluster size. We discuss this issue further in Section 3.2. George and Bowman (1995) and Bowman and George (1995) investigate maximum likelihood estimation for one particular parameterization of the saturated model. However, they note that with unequal cluster sizes, their estimates are not necessarily admissible because they do not satisfy the built-in non-negativity constraints for joint cell probabilities.

In this paper we argue that, as a device to obtain feasible MLEs, groups of different sizes may be viewed as coming from a sample of equal size clusters with some observations missing at random, even if this is not necessarily the mechanism whereby the unequal cluster sizes arise. Then we propose the EM algorithm for full likelihood estimation of the saturated model parameters. The resulting MLEs are always admissible because the model constraints are automatically accounted for in the likelihood maximization. The paper is structured as follows: Section 2 examines parameterizations of models for

clustered exchangeable binary variables; estimation procedures are developed in Section 3 and one technique of testing the model assumptions is presented in Section 4. Section 5 describes an application to teratological studies, as well as implications for the design of the PRECISE trial. Several directions for future research are outlined in a concluding section.

2. Exchangeable Binary Distributions

Let $(Y_1, \dots, Y_r)$ be a cluster of exchangeable binary random variables, where $Y_i = 0$ (“failure”) or 1 (“success”), $i = 1, \dots, r$. There are different ways of parameterizing the joint distribution and here we discuss four of them.

The Bahadur representation (Bahadur (1961)) uses the marginal response probability $p = \Pr(Y_i = 1)$ and correlations of all orders $(\rho_2, \rho_3, \dots, \rho_r)$, where $\rho_k = E[(Y_1 - p) \dots (Y_k - p)] / \{[p(1 - p)]^{k/2}\}$. By exchangeability, ρ_k is the same for any subset of size k of the r responses $(Y_1, \dots, Y_r)$. This parameterization was also used by Prentice (1988). In particular, the pairwise correlation is

$$\rho = \rho_2 = \frac{E(Y_i Y_j) - p^2}{p(1 - p)}. \quad (1)$$

There are restrictions on the range of possible values for $(p, \rho = \rho_2, \dots, \rho_r)$. For example, Figure 1 gives the feasible regions for (p, ρ) when $r = 2, 3, 4, \infty$.

[Figure 1 about here.]

The response conditional parameterization is in terms of probabilities $\gamma_0, \dots, \gamma_{r-1}$, where

$$\gamma_s = \Pr(Y_i = 1 \mid \sum_{j=1, j \neq i}^r Y_j = s), \quad i = 1, \dots, r, \quad s = 0, \dots, r-1.$$

This parameterization can lead to a class of *response conditional models* (Pendergast et al., 1996) and is used for example by Qu et al. (1987), and Connolly and Liang (1988).

The cluster sum distribution representation uses probabilities $q_0, \dots, q_r$ with $\sum_{s=0}^r q_s = 1$, where $q_s = \Pr(S = s)$, $s = 0, \dots, r$, and $S = \sum_{i=1}^r Y_i$. Note that, in terms of the Bahadur representation, we have

$$p = \sum_{s=1}^r \Pr(Y_i = 1 \mid S = s) \Pr(S = s) = \frac{1}{r} \sum_{s=0}^r s q_s. \quad (2)$$

Also, for $i \neq j$, since

$$E(Y_i Y_j) = \sum_{s=2}^r \Pr(Y_i = 1, Y_j = 1 \mid S = s) \Pr(S = s) = \sum_{s=2}^r \frac{s(s-1)}{r(r-1)} q_s, \quad (3)$$

the correlation ρ may be obtained as a function of $\{q_s\}$ from (1) and (2), etc.

The Bowman–George representation (Bowman and George (1995)) uses joint success probabilities $1 = \lambda_0 \geq \lambda_1 \geq \dots \geq \lambda_r$, where

$$\lambda_k = \Pr(Y_1 = 1, Y_2 = 1, \dots, Y_k = 1), \quad k = 1, \dots, r.$$

Then $p = \lambda_1$, and the correlations are (Bowman and George, 1995, Eqn. (7))

$$\rho_k = \frac{\sum_{j=0}^k (-1)^{k-j} \binom{k}{j} \lambda_1^{k-j} \lambda_j}{[\lambda_1(1 - \lambda_1)]^{k/2}}. \quad (4)$$

In particular,

$$\rho = \frac{\lambda_2 - \lambda_1^2}{\lambda_1(1 - \lambda_1)}. \quad (5)$$

The parameters need to satisfy the constraints (George and Bowman, 1995, Eqn. 3.3):

$$\Delta^{r-s}(\lambda_s) \geq 0, \quad s = 0, \dots, r, \quad (6)$$

where forward differences $\Delta(\lambda_s) = \lambda_s - \lambda_{s+1}$, $\Delta^2(\lambda_s) = \Delta(\lambda_s) - \Delta(\lambda_{s+1})$ etc., in order that all the cluster sum probabilities be non-negative — see (7) below. The feasible regions for (λ_1, λ_2) when $r = 2, 3, 4, \infty$ are given in Figure 2, and correspond to those for (p, ρ) shown in Figure 1.

[Figure 2 about here.]

Of course, the four parametric characterizations are equivalent and one can easily alternate between them. For example, we have (4) relating $\{\lambda_k\}$ to $\{\rho_k\}$. Also

$$q_s = \binom{r}{s} \sum_{j=0}^{r-s} (-1)^j \binom{r-s}{j} \lambda_{s+j}, \quad \text{for } s = 0, \dots, r, \quad (7)$$

(see Bowman and George (1995, Eqn 5)), and

$$\gamma_s = \frac{(s+1) \cdot q_{s+1}}{(s+1) \cdot q_{s+1} + (r-s) \cdot q_s}, \quad \text{for } s = 0, \dots, r-1.$$

The most general model is the saturated model in which the parameter sets, either $(p, \rho = \rho_2, \rho_3, \dots, \rho_r)$, $(\gamma_0, \gamma_1, \dots, \gamma_{r-1})$, $(\lambda_1, \dots, \lambda_r)$ or $(q_0, \dots, q_r)$ lie in an r -dimensional space. The saturated model places no particular structure on the joint distribution of $(Y_1, \dots, Y_r)$. More parsimonious (unsaturated) models have been proposed, where the r defining parameters are expressed as functions of a smaller collection of parameters $\theta_1, \dots, \theta_{r'}$, say, where $r' < r$. Such special cases include the binomial for $r' = 1$, the beta-binomial (e.g. Rosner (1984), Prentice (1986)), Altham's (1978) multiplicative generalization of the binomial and the correlated binomial (Kupper and Haseman (1978), Altham (1978)) for $r' = 2$, the beta correlated binomial (Paul, 1987), and the Polya-Eggenberger distribution

(Feller (1968) Eqn. V.2.3) for $r' = 3$. For general $r' < r$, one can consider binomial mixture models (e.g. Brooks et al. (1997)), or modified binomial distributions (Ng (1989)). The saturated model may be used to test the goodness of fit of such more parsimonious models, as in our example in Section 5.1.

3. Maximum Likelihood Estimation

3.1 Equal Cluster Size

Suppose we observe K i.i.d. clusters of size r , with the responses of the k th cluster denoted by $Y_{k1}, \dots, Y_{kr}$ for $1 \leq k \leq K$. Let n_s be the number of clusters containing exactly s Y -values equal to one ($0 \leq s \leq r$). Thus $n_0 + \dots + n_r = K$ and the likelihood is simply proportional to the multinomial likelihood $\prod_{s=0}^r q_s^{n_s}$. For unsaturated models, the parametric form $q_s(\theta_1, \dots, \theta_{r'})$ is inserted into this likelihood which is maximized with respect to $(\theta_1, \dots, \theta_{r'})$. Usually, numerical methods are required (e.g. Rosner (1984), George and Bowman (1995)).

For the saturated model it is easiest to work with the cluster sum probabilities $\{q_s\}$. The MLEs are simply given by $\hat{q}_s = n_s/K$. Substituting those in (2), we obtain the MLE of the marginal probability p as

$$\hat{p} = \frac{1}{rK} \sum_{s=0}^r s n_s = \frac{1}{rK} \sum_{k=1}^K \sum_{i=1}^r Y_{ki} = \bar{Y}, \quad (8)$$

which is of course the intuitive estimator, coinciding with the method of moments estimator. From (8) we have

$$\text{Var}(\hat{p}) = \frac{1}{Kr} p(1-p)[1 + (r-1)\rho]. \quad (9)$$

Note this implies the restriction $-1/(r-1) \leq \rho \leq 1$ — see Figure 1. Also, from (3), we obtain

$$\hat{E}(Y_i Y_j) = \sum_{s=2}^r \frac{s(s-1)}{r(r-1)} \cdot \frac{n_s}{K}.$$

Hence the MLE of the intracluster correlation is $\hat{\rho} = \frac{\hat{E}(Y_i Y_j) - \hat{p}^2}{\hat{p}(1-\hat{p})}$. George and Bowman (1995) work with the $\{\lambda_s\}$ and explicit expressions for the MLEs $\{\hat{\lambda}_s\}$ in the equal cluster size case are given by their Equation (12).

3.2 Clusters of Varying Sizes

Suppose now the K clusters are independent but of varying sizes, with the k -th cluster having r_k observations denoted $Y_{k1}, \dots, Y_{kr_k}$, for $1 \leq k \leq K$. Let the maximum cluster size be $R = \max\{r_k; 1 \leq k \leq K\}$, and let $K_r = \sum_{k=1}^K I(r_k = r)$ be the number of clusters of size r , hence $K = \sum_{r=1}^R K_r$. Also, let n_{sr} denote the total number of clusters of size r with exactly s successes.

If we wish to make inferences based on combined information from clusters of varying sizes, it makes sense to use model formulations in which the parameters have the same meaning irrespective of the cluster size. We term this the “interpretability” assumption. For example, in the Bahadur representation, the marginal mean p and correlation ρ_k are required to be the same for all clusters of size $r \geq k$. Equivalently, in the Bowman–George representation λ_k is required to be the same for all clusters of size $r \geq k$. Note that it makes no sense to require this of the parameter sets $(q_0, q_1, \dots, q_k)$ or $(\gamma_0, \gamma_1, \dots, \gamma_{k-1})$, which naturally must depend on the cluster size $r \geq k$ from their definitions.

This assumption needs to be justified in any given application. For exam-

ple to the contrary, it could be that the marginal mean p *does* depend on the cluster size r . This might be true in teratological studies where survival of offspring might depend on the size of the litter, if there was a competition (or cooperation) effect. However, in other situations such as familial aggregation studies of disease or group randomized trials, the interpretability assumption may be reasonable. Without it, it is difficult to combine the information from clusters of different size without imposing some model assumption for the effect of cluster size on the joint distribution. Some unsaturated models, such as the binomial, beta-binomial and Polya-Eggenberger with nonnegative pairwise correlation, do satisfy the interpretability assumption, whereas others, such as Altham's multiplicative generalization of the binomial, do not. In any case, the interpretability assumption can be tested – see Section 4.

From now on we make the interpretability assumption. The full likelihood is a product of up to R multinomial likelihoods of the form considered in Section 3.1 — one factor corresponding to each value of r with $K_r \neq 0$. Because unlike the $\{q_k\}$, the $\{\lambda_k\}$ do not depend on r , we rewrite each q_s in terms of $\lambda_1, \dots, \lambda_R$ using (7). The likelihood becomes

$$L = \prod_{r=1}^R \prod_{s=0}^r \left[\binom{r}{s} \sum_{j=0}^{r-s} (-1)^j \binom{r-s}{j} \lambda_{s+j} \right]^{n_{sr}}. \quad (10)$$

This likelihood could then be maximized directly with respect to $\lambda_1, \dots, \lambda_R$ using numerical methods. However, it is difficult to incorporate the feasibility constraints (6).

Bowman and George (1995, Eqn. (15)) proposed the estimators

$$\tilde{\lambda}_k = \frac{1}{K} \sum_{i=k}^R \sum_{j=0}^{i-k} \frac{\binom{i-k}{j}}{\binom{i}{j}} \cdot n_{i,i-j}.$$

However, they remark (p.874) that these estimates do not necessarily satisfy the constraints (6), as we noted in Section 1. For example, if the sample consists of two clusters with $K_1 = 1$ and $K_2 = 1$ and responses $\{1\}$ and $\{1, 0\}$, then $\tilde{\lambda}_1 = \tilde{p} = .75$ and $\tilde{\lambda}_2 = 0$. This leads to the inadmissible values $\tilde{\rho} = \frac{\tilde{\lambda}_2 - \tilde{\lambda}_1^2}{\tilde{\lambda}_1(1 - \tilde{\lambda}_1)} = -3$ and $\tilde{q}_0 = -.5$, $\tilde{q}_1 = 1.5$, $\tilde{q}_2 = 0$; whereas the MLEs are $\hat{\lambda}_1 = \hat{p} = .5$, $\hat{\lambda}_2 = 0$, $\hat{\rho} = -1$ and $\hat{q}_0 = 0$, $\hat{q}_1 = 1$, $\hat{q}_2 = 0$.

In order to alleviate these difficulties we shall show that the likelihood arising from a situation where all data come from clusters of equal size R , but clusters with sizes $r < R$ have $R - r$ observations missing completely at random (MCAR) is the same as that given by (10). Let us denote by Pr_r probability statements conditional on the cluster size being r and let $q_{s,r} = \text{Pr}_r(\sum_{i=0}^r Y_i = s)$. By the interpretability assumption, each λ_k does not depend on r . For a generic cluster of size r with observations $(y_1, \dots, y_r)$ and cluster sum $\sum_{i=0}^r y_i = s$, we have

$$\text{Pr}_r(y_1, \dots, y_r) = q_{s,r} / \binom{r}{s} \quad (11)$$

$$= \sum_{j=0}^{r-s} (-1)^j \binom{r-s}{j} \lambda_{s+j} \quad (12)$$

$$= \sum_{t=s}^{R-r+s} \binom{R-r}{t-s} \sum_{j=0}^{R-t} (-1)^j \binom{R-t}{j} \lambda_{t+j} \quad (13)$$

$$= \sum_{t=s}^{R-r+s} \binom{R-r}{t-s} \cdot q_{t,R} / \binom{R}{t}, \quad (14)$$

which, by exchangeability, is the probability of observing $(y_1, \dots, y_r)$ in a cluster of size R , when responses $y_{r+1}, \dots, y_R$ are MCAR. Here, (11) follows from exchangeability, (12) and (14) from Bowman and George (1995), and (13) can be proven by induction on $R - r$, for any integers $s, r > 0$ and $R \geq r$.

Hence, under the interpretability assumption, the likelihood of any particular cluster of size r is the same as the likelihood of those outcomes arising from a larger cluster of size R but with $R - r$ observations missing completely at random. Thus it is natural to consider use of the EM algorithm (Dempster, Laird and Rubin, 1977) to obtain maximum likelihood estimates.

For any cluster k with $r < R$ observations, we treat $S = Y_{k1} + \dots + Y_{kr}$ as the “incomplete data” and $T = S + Y_{k,r+1} + \dots + Y_{kR}$ as the “complete data”. We set

$$\begin{aligned} p_{rst} &= \Pr(T = t \mid S = s) = \frac{\Pr(S = s \mid T = t) \Pr(T = t)}{\sum_{t'} \Pr(S = s \mid T = t') \Pr(T = t')} \\ &= \frac{\left[\binom{t}{s} \binom{R-t}{r-s} / \binom{R}{r} \right] q_t}{\sum_{t'=s}^{R-r+s} \left[\binom{t'}{s} \binom{R-t'}{r-s} / \binom{R}{r} \right] q_{t'}} = \frac{\left[\binom{t}{s} \binom{R-t}{r-s} \right] q_t}{\sum_{t'=s}^{R-r+s} \left[\binom{t'}{s} \binom{R-t'}{r-s} \right] q_{t'}} \end{aligned}$$

for $t = s, s+1, \dots, R-r+s$, where we have used the exchangeability of the Y 's. Also, let $p_{Rst} = 1$ if $s = t$ and 0 otherwise. The likelihood is then

$$L = \prod_{r=1}^R \prod_{k=1}^{K_r} \Pr_r(y_{k1}, \dots, y_{kr}) = \prod_{r=1}^R \prod_{s=0}^r \left[\sum_{t=s}^{R-r+s} q_t \cdot \frac{\binom{R-r}{t-s}}{\binom{R}{t}} \right]^{n_{sr}}.$$

The EM algorithm proceeds as follows. We start with initial estimates $q_0^{(0)}, q_1^{(0)}, \dots, q_R^{(0)}$, say. For example, take $q_0^{(0)} = q_1^{(0)} = \dots = q_R^{(0)} = 1/(R+1)$. At the E-step, we estimate the number of clusters with exactly t successes if the “missing data” were present so that all clusters have common size R . Call this quantity n'_t ; it is given by

$$n'_t = \sum_{r=1}^R \sum_{s=\max(0, t+r-R)}^{\min(t, r)} n_{sr} p_{rst}.$$

Note these are not necessarily integers. At the M-step, we update the estimates of $q_0, \dots, q_R$ using the complete data. Since the complete data consist

of clusters of common size R , we can use the methods of Section 3.1. For unsaturated models, we use the frequencies n'_t ; ($0 \leq t \leq R$) to estimate the parameters $\theta_1, \dots, \theta_{r'}$ and hence the parameters $q_0, \dots, q_R$. For the saturated model, the updated estimate of q_t is n'_t/K , $0 \leq t \leq R$. Upon convergence,

$$\hat{p} = \frac{1}{RK} \sum_{t=1}^R t n'_t, \quad \hat{\rho} = \frac{\hat{E}(Y_1 Y_2) - \hat{p}^2}{\hat{p}(1 - \hat{p})}, \quad \hat{E}(Y_1 Y_2) = \sum_{t=2}^R \frac{t(t-1)}{R(R-1)} \cdot \frac{n'_t}{K}. \quad (15)$$

Since $\hat{p} = \frac{1}{R} \sum_{t=1}^R t \hat{q}_t$, it follows that

$$\text{Var}(\hat{p}) = \frac{1}{R^2} \left[\sum_{t=1}^R t^2 \cdot \text{Var}(\hat{q}_t) + 2 \sum_{t_1 < t_2} t_1 t_2 \cdot \text{cov}(\hat{q}_{t_1}, \hat{q}_{t_2}) \right].$$

Let Σ_q be the variance–covariance matrix of the vector of estimates $\hat{\mathbf{q}} = (\hat{q}_1, \hat{q}_2, \dots, \hat{q}_R)$. A formula for Σ_q in the general multivariate case is given in Meng and Rubin (1991):

$$\Sigma_q = \mathbf{I}_{oc}^{-1} + \mathbf{I}_{oc}^{-1} \cdot (\mathbf{I} - \mathbf{DM})^{-1} \mathbf{DM}^{-1},$$

where $\mathbf{I}$ is the $R \times R$ identity matrix, and $\mathbf{I}_{oc} = E(\mathbf{I}_o(\mathbf{q}|\mathbf{Y})|\mathbf{Y}_{obs}, \mathbf{q})|_{\mathbf{q}=\hat{\mathbf{q}}}$ is the expectation of the complete–data observed information matrix $\mathbf{I}_o(\mathbf{q}|\mathbf{Y}) = -\frac{\partial^2 \log f(\mathbf{Y}|\mathbf{q})}{\partial \mathbf{q}^2}$ over the conditional distribution $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{q})$, evaluated at $\mathbf{q} = \hat{\mathbf{q}}$. Also, $\mathbf{DM} = \left(\frac{\partial M_j(\mathbf{q})}{\partial q_i} \right) \Big|_{\mathbf{q}=\hat{\mathbf{q}}}$ is the $R \times R$ Jacobian matrix for $\mathbf{M}(\mathbf{q}) = (M_1(\mathbf{q}), \dots, M_R(\mathbf{q}))$ evaluated at $\mathbf{q} = \hat{\mathbf{q}}$, where $\mathbf{q} \rightarrow \mathbf{M}(\mathbf{q})$ is the mapping from the parameter space onto itself implicitly defined by the EM algorithm such that $\mathbf{q}^{(i+1)} = \mathbf{M}(\mathbf{q}^{(i)})$ at the i –th iteration of the algorithm.

In the case of the multinomial likelihood we have

$$\mathbf{I}_{oc} = \frac{K}{1 - \sum_{t=1}^R \hat{q}_t} \cdot \mathbf{1}_R + K \cdot (\text{diag}(\hat{q}_1, \dots, \hat{q}_R))^{-1},$$

where $\mathbf{1}_R$ is the $R \times R$ matrix of ones and $\mathbf{diag}(\hat{q}_1, \dots, \hat{q}_R)$ is the diagonal matrix with elements $\hat{q}_1, \dots, \hat{q}_R$. Also, for $j = 1, \dots, R$

$$\mathbf{DM}_{jj} = \frac{1}{K} \cdot \sum_{r=1}^R \sum_{s=\max(0, j+r-R)}^{\min(j, r)} n_{sr} \binom{j}{s} \binom{R-j}{r-s} \cdot \frac{\sum_{t'=s, t' \neq j}^{R-r+s} \binom{t'}{s} \binom{R-t'}{r-s} \hat{q}_{t'}}{\left[\sum_{t'=s}^{R-r+s} \binom{t'}{s} \binom{R-t'}{r-s} \hat{q}_{t'} \right]^2},$$

and for $i, j = 1, \dots, R, \ i \neq j$,

$$\mathbf{DM}_{ji} = -\frac{1}{K} \cdot \sum_{r=1}^R \sum_{s=\max(0, j+r-R, i+r-R)}^{\min(i, j, r)} n_{sr} \cdot \frac{\binom{j}{s} \binom{R-j}{r-s} \binom{i}{s} \binom{R-i}{r-s} \hat{q}_j}{\left[\sum_{t'=s}^{R-r+s} \binom{t'}{s} \binom{R-t'}{r-s} \hat{q}_{t'} \right]^2}.$$

Since $p_{11} = \frac{1}{R(R-1)} \sum_{t=2}^R t(t-1)\hat{q}_t$, we obtain

$$\hat{\rho} = g(\hat{\mathbf{q}}) = \frac{\frac{1}{R(R-1)} \sum_{t=2}^R t(t-1)\hat{q}_t - \left[\frac{1}{R} \sum_{t=1}^R t\hat{q}_t \right]^2}{\left[\frac{1}{R} \sum_{t=1}^R t\hat{q}_t \right] \cdot \left[1 - \frac{1}{R} \sum_{t=1}^R t\hat{q}_t \right]}.$$

Using the delta method, the asymptotic variance of $\hat{\rho}$ is

$$\text{Var}(\hat{\rho}) = \nabla g(\hat{\mathbf{q}}) \cdot \Sigma_q \cdot (\nabla g(\hat{\mathbf{q}}))',$$

where $\nabla g(\hat{\mathbf{q}}) = (\nabla g(\hat{\mathbf{q}})_i)_{i=1, \dots, R}$ and

$$(\nabla g(\hat{\mathbf{q}}))_i = \frac{i}{R(R-1) \cdot \hat{p}^2(1-\hat{p})^2} \cdot \left[\hat{p}(i(1-\hat{p}) - R\hat{p}) - \frac{1-2\hat{p}}{R} \cdot \sum_{t=2}^R t^2 \hat{q}_t \right].$$

4. Testing Model Assumptions

The most important feature of the interpretability assumption is the constant probability of success for all individuals irrespective of their cluster size. This assumption is not always reasonable, and Armitage's trend test (Armitage, 1955) may be used to check whether the success probability varies with the

cluster size. However, Armitage's test was designed for observations that are independent, and it can give an erroneously inflated level of significance when applied to correlated data. Here we generalize Armitage's test to account for intracluster correlation when the cluster size is variable, following the approach of Le (1988) who considered trends with a covariate in paired binary observations.

With the notation of Section 3.2, let $a_r = \sum_{s=1}^r s n_{sr}$ for $r = 1, \dots, R$ be the total number of successes in clusters of size r . Also, let x_r be a score assigned to clusters of size r in Armitage's test. A common choice is $x_1 = -\frac{1}{2}(R-1), x_2 = -\frac{1}{2}(R-3), \dots, x_R = \frac{1}{2}(R-1)$. The estimated marginal probability p_r in clusters of size r is $\hat{p}_r = a_r / r K_r$, for $r = 1, \dots, R$, and the overall estimate of the marginal probability is $\hat{p}$ from (15). Consider the test statistic

$$T = \sum_{r=1}^R r K_r x_r \hat{p}_r.$$

Then

$$\hat{E}_0(T) = \hat{p} \sum_{r=1}^R r K_r x_r$$

and

$$\widehat{\text{Var}}_0(T) = \hat{p}(1 - \hat{p}) \sum_{r=1}^R r K_r x_r^2 [1 + (r - 1)\hat{\rho}]$$

are the estimated mean and variance of T under the null hypothesis of no trend. Here $\hat{\rho}$ is the estimated correlation, common to all clusters. Let

$$X^2 = \frac{[T - \hat{E}_0(T)]^2}{\widehat{\text{Var}}_0(T)}.$$

Under the null hypothesis, X^2 has an asymptotic χ_1^2 distribution. Thus the p -value of the correlated trend test is $\Pr(X^2 \geq \chi_1^2)$.

5. Applications

5.1 *Chronic Obstructive Pulmonary Disease*

The possible familial aggregation of a specific disease is a major concern of genetic epidemiologists, as it may indicate potential genetic and environmental factors contributing to disease etiology (Khoury et al. (1993)). The exchangeable models of Section 2 find a natural application in studies of familial aggregation. The intraclass correlation ρ may be estimated as a measure of familial association; then the odds ratio can be computed from

$$\Psi = 1 + \frac{\rho}{(1 - \rho)^2} \cdot \left[\frac{1}{p} + \frac{1}{1 - p} \right] = \frac{\lambda_2(1 - 2\lambda_1 + \lambda_2)}{(\lambda_1 - \lambda_2)^2}.$$

As an example, we analyze the subset of the data from a case-control study of familial aggregation of chronic obstructive pulmonary disease (COPD) (Cohen (1980)), which was analyzed by Liang et al. (1992) and presented in their Table 5. The main objective was to assess the presence and magnitude of familial aggregation of impaired pulmonary function (IPF) due to genetic factors. The binary response indicates whether each sibling of a COPD patient had IPF. A cluster in this analysis is a set of siblings, and the data set contains 203 siblings from 100 clusters with sizes varying from 1 to 6. We fit three models to the data — the binomial, the beta-binomial and the saturated model of Section 2. The maximum likelihood estimates for the saturated model were computed using the EM algorithm of Section 3.2, while a Newton-Raphson maximization routine was used for the other two models. The results of our analysis are summarized in Table 1.

[Table 1 about here.]

Following Brooks et al. (1997), we compare the three models using the deviance (dev), the Akaike (1973) and Bayesian (Schwarz (1978)) information criteria. Here, $\text{dev} = -2\log(L)$, $\text{AIC} = -2\log(L) + 2k$, and $\text{BIC} = -2\log(L) + k\log(K)$, where L is the likelihood, k is the number of estimated parameters, and K is the total number of clusters. Conclusions from the different information criteria are in broad agreement. Examination of Table 1 leads us to conclude that the binomial is not an adequate model, whereas the beta-binomial does fit the data well. Indeed it would be difficult to improve on the beta-binomial fit, as comparison with the saturated log-likelihood shows.

The MLEs of p from the three models differ little. The estimated $\hat{p} = 0.200$ in the saturated model indicates evidence of familial aggregation of disease. Also, note that the MLEs $\hat{p} = 0.284$ and $\hat{p} = 0.200$ from the saturated model lead to an estimated odds-ratio of $\hat{\Psi} = 2.54$. Using GEE techniques, Liang et al. (1992, Table 7) fitted a marginal logistic regression model and report an estimated odds-ratio of $2.69 (= \exp(0.99))$ when covariates are also not included in the model.

The saturated model relies on the assumption that the marginal probability of success is the same in all clusters. This assumption needs to be tested since, for example, the empirical raw response rate in clusters of size 6 is approximately twice the rate in clusters of size 1. It seems possible that the probability of disease increases with cluster size. Indeed, with a p -value of 0.0451 the standard Armitage trend test rejects the null hypothesis of no trend at 5% significance level. However, using the correction for intracluster correlation proposed in Section 4, the p -value of this modified trend test be-

comes 0.088 corresponding to $\hat{\rho} = 0.2$. Thus the null hypothesis of no trend cannot be rejected. In fact, the p -value of the modified test increases with ρ , and Table 2 presents the results of a sensitivity analysis.

[Table 2 about here.]

5.2 *Implications for Statistical Design: The Proposed PRECISE Trial*

We return to the sample size problem for the PRECISE trial as described in Section 1. When testing equality of success parameters in two samples of clusters of size r , it follows from (9) that the sample size needed has to be inflated by a factor of $1 + (r - 1)\rho$ over the number that would be required were all observations independent. This is called the design effect (Scott and Holt (1982), Neuhaus and Segal (1993)).

For the PRECISE trial, let us suppose that clusters are of size at most two, and consider the worst case when all observations come in pairs ($r = 2$). From historical records it was estimated that the 5-year cancer incidence rate for the control group in this population would be approximately $p = 0.075$. The Principal Investigator also was willing to estimate that the probability that a second member of the household pair would become diseased in the 5-year period would increase from p to ξp with ξ approximately 1.04 upon knowledge that the other member was an incident case, although this was less certain. This specification leads to a preliminary estimate $\rho = \frac{\xi p - p}{1 - p} = 0.0032$ and the effect on sample size is negligible. A sensitivity analysis presented in Table 3 demonstrates that this remains true for a range of p and ξ values around the specified values. In view of all of its advantages as outlined in Section 1, the investigators decided to proceed with the clustered design.

[Table 3 about here.]

6. Discussion and Future Research

Saturated models are useful tools for analyzing clustered exchangeable binary data when higher order interactions are of interest, and also as a benchmark with which to compare more parsimonious models. Likelihood methods can serve as the basis for inference. The EM algorithm facilitates computation, and is especially valuable when clusters are of varying sizes. We have investigated a large number of examples and our experience has been that the EM algorithm has converged rapidly. This happens even in examples with maximum cluster size of the order of $R = 150$, which might arise in community randomized nutrition trials — for example, the vitamin A supplementation trial in Nepal (West et al., 1995).

The EM methodology can be extended to cover the situation where the binary outcome variables are exchangeable only after adjustment for covariates, or where the outcome scale is other than binary — categorical or survival endpoints, for example. These and other generalizations are the subject of further research.

REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Czaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Budapest: Akademia Kaido.
- Altham, P. (1978). Two generalizations of the binomial distribution. *Applied Statistics* **27**, 162–167.
- Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.
- Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In Solomon, H., editor, *Studies in Item Analysis and Prediction*, volume VI of *Stanford Mathematical Studies in the Social Sciences*. Stanford, California: Stanford University Press.
- 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.
- Brooks, S. P., Morgan, B. J. T., Ridout, M. S. and Pack, S. E. (1997). Finite mixture models for proportions. *Biometrics* **53**, 1097–1115.
- Cohen, B. H. (1980). Chronic obstructive pulmonary disease: a challenge in genetic epidemiology. *American Journal of Epidemiology* **112**, 274–288.
- Connolly, M. and Liang, K. Y. (1988). Conditional logistic regression models for correlated binary data. *Biometrika* **75**, 501–506.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of*

- Royal Statistical Society B* **39**, 1–38.
- Feller, W. B. (1968). *An Introduction to Probability Theory and its Applications*. Wiley, New York, third edition.
- George, E. O. and Bowman, D. (1995). A full likelihood procedure for analysing exchangeable binary data. *Biometrics* **51**, 512–523.
- Khoury, M. J., Beaty, T. H. and Cohen, B. H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press, New York.
- Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **34**, 69–76.
- Le, C. T. (1988). Testing for linear trends in proportions using correlated otolaryngology or ophthalmology data. *Biometrics* **44**, 291–303.
- Liang, K. Y., Qaqish, B. and Zeger, S. (1992). Multivariate regression analyses for categorical data. *Journal of Royal Statistical Society B* **54**, 3–40.
- Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.
- Natarajan, R., Turnbull, B. W., Slate, E. H. and Clark, L. C. (1996). A computer program for sample size and power calculations in the design of multi–arm and factorial clinical trials with survival time endpoints. *Computer Methods and Programs in Biomedicine* **49**, 137–147.
- Neuhaus, J. M. and Segal, M. R. (1993). Design effects for binary regression models fitted to dependent data. *Statistics in Medicine* **12**, 1259–1268.
- Ng, T. H. (1989). A new class of modified binomial distributions with applications to certain toxicological experiments. *Communications in Statistics – Theory and Methods* **18**, 3477–3492.

- Paul, S. R. (1987). On the beta-correlated binomial (BCB) distribution — a three-parameter generalization of the binomial distribution. *Communications in Statistics – Theory and Methods* **16**, 1473–1478.
- Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M. and Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review* **64**, 89–118.
- 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* **394**, 321–327.
- Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.
- Qu, Y., Williams, G. W., Beck, G. J. and Goormastic, M. (1987). A generalized model of logistic regression for clustered data. *Communications in Statistics* **16**, 3447–3476.
- Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired-data situations. *Biometrics* **40**, 1025–1035.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.
- Scott, A. J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* **77**, 848–854.
- West, K. Katz, J. et al. (1995). Mortality of infants under 6 months of age supplemented with vitamin A: a randomized, double-masked trial in Nepal. *American Journal of Clinical Nutrition* **62**, 143–148.

Figure 1. Feasible values of (p, ρ) for exchangeable binary distributions and selected values of model dimension $r = 2, 3, 4, \infty$. The upper bound on ρ for all feasible regions is $\rho = 1$. Shaded areas are infeasible, with darker shading corresponding to the smaller values of r . Each feasible region is nested within the previous one for increasing r . In the limiting case $r \rightarrow \infty$, we must have $0 \leq \rho \leq 1$, for all p .

Figure 2. Feasible values of (λ_1, λ_2) for exchangeable binary distributions, corresponding to constraints (6) and selected values of model dimension $r = 2, 3, 4, \infty$. The first diagonal gives the upper bound of all feasible regions ($\lambda_2 \leq \lambda_1$). Shaded areas are infeasible, with darker shading corresponding to the smaller values of r . Each feasible region is nested within the previous one for increasing r . In the limiting case $r \rightarrow \infty$, we must have $\lambda_1^2 \leq \lambda_2 \leq \lambda_1$.

Table 1
Maximum likelihood estimation for COPD data and comparison of models.

	Binomial	Beta-binomial	Saturated model
Number of parameters	1	2	6
$\hat{p}$	0.296 [0.032]	0.282 [0.036]	0.284 [0.038]
$\hat{\rho}$	–	0.270 [0.145]	0.200 [0.089]
$-\log L$	93.033	88.620	87.050
dev	186.065	177.239	174.099
AIC	188.065	181.239	186.099
BIC	190.671	186.449	201.730

Table 2

Sensitivity analysis of the modified trend test. Table reports the p -values for COPD data, at varying levels of intraclass correlation. Estimated correlation is $\hat{\rho} = 0.2$.

ρ	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
p -value	.045	.066	.088	.110	.131	.151	.170	.189	.206	.223

Table 3

Sensitivity analysis. Table shows values of ρ for various combinations of ξ and p . With clusters of size $r = 2$, the usual sample size requirement for independent observations must be inflated by a factor of $(1 + \rho)$.

ξ	1.00	1.02	1.04	1.06	1.08	1.10
p						
0.015	0.000	0.0003	0.0006	0.0009	0.0012	0.0015
0.020	0.000	0.0004	0.0008	0.0012	0.0016	0.0020
0.030	0.000	0.0006	0.0012	0.0019	0.0025	0.0031
0.040	0.000	0.0008	0.0017	0.0025	0.0033	0.0042
0.050	0.000	0.0011	0.0021	0.0032	0.0042	0.0053
0.060	0.000	0.0013	0.0026	0.0038	0.0051	0.0064
0.070	0.000	0.0015	0.0030	0.0045	0.0060	0.0075
0.080	0.000	0.0017	0.0035	0.0052	0.0070	0.0087
0.085	0.000	0.0019	0.0037	0.0056	0.0074	0.0093