International Statistical Review (1991), **59**, 1, pp. 25–35. Printed in Great Britain

A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data

J.M. Neuhaus[1], J.D. Kalbfleisch[2] and W.W. Hauck[1]

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94143-0560, USA. [2]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

Summary

Clustered or correlated samples of binary responses arise frequently in practice due to repeated measurements or to subsampling the primary sampling units. Several recent approaches address intracluster correlation in binary regression problems including cluster-specific methods such as those based on mixed-effects logistic models and population-averaged methods such as those based on beta-binomial models. This paper considers the interpretations of the regression parameters in these two general approaches. We show that, unlike models for correlated Gaussian outcomes, the parameters of the cluster-specific and population-averaged models for correlated binary data describe different types of effects of the covariates on the response probabilities. In the case of random intercepts, we show that the covariate effects measured by the population-averaged approach are closer to zero than those of the cluster-specific approach when the cluster-specific model holds and that the difference in the magnitude of the covariate effects is increasing with intra-cluster correlation. The case of random slopes is also examined. These results are valid for arbitrary random effects distributions and are demonstrated using data on the ability to obtain samples of breast fluid from women.

Key words: Binary data; Clustered data; Intraclass correlation; Mixed-effects models; Population-averaged models.

1 Introduction

Clustered or correlated samples of binary data arise frequently in many fields of application. This clustering may be due to repeated measurements of individuals over time, as in longitudinal studies, or may be due to subsampling the primary sampling units. Examples of the latter are common, for example, in ophthalmology where two eyes form a cluster but observations are taken on each eye, and in periodontology where the mouth is the cluster but the data are gathered from multiple sites within the mouth. It is well known that observations in the same cluster tend to exhibit intracluster correlation, and that standard methods of analysis that ignore the clusters tend to be inadequate; models fit poorly and variances are poorly estimated. See, for example, Altham (1978) and Kupper et al. (1986). Several approaches have recently been proposed to analyse binary data gathered in clusters. See, for example, Ware, Lipsitz & Speizer (1988) and Prentice (1988) for reviews and discussion. Most of these approaches can be grouped into two classes described by Zeger, Liang & Albert (1988): cluster-specific and population-averaged.

To be more specific, suppose that data are collected on a binary outcome, Y_{ij}, together with a vector of covariates, $\mathbf{X}_{ij}$. The data are gathered in clusters or groups and $i = 1, \ldots, m$ indexes clusters while $j = 1, \ldots, n_i$ indexes units within clusters. With the cluster-specific approach, the probability distribution of Y_{ij} is modelled as a function of the covariates, $\mathbf{X}_{ij}$, and parameters α_i specific to the ith cluster. Examples of this approach include the mixed-effects logistic model, e.g. Stiratelli, Laird & Ware (1984) and Anderson & Aitkin (1985), and the conditional likelihood approach for matched pair data, e.g. Breslow & Day (1980). With the population-averaged approach, the marginal or population-averaged expectation of Y_{ij} is modelled as a function of the covariates $\mathbf{X}_{ij}$. Examples of this approach include the generalized estimating equation (GEE) approach of Liang & Zeger (1986) and the beta-binomial regression model (Williams, 1975; Prentice, 1986).

To introduce these two approaches, consider the linear mixed-effects model,

$$Y_{ij} \mid \alpha_i = \mu + \alpha_i + \beta X_{ij} + \varepsilon_{ij}, \tag{1}$$

where, conditional on α_i, the ε_{ij} are independent $N(0, \sigma_e^2)$ variates and the cluster-specific effects α_i are independent $N(0, \sigma_A^2)$ variates. In the model (1), the parameters β measure a cluster-specific effect of the covariates X. This same model has a simple interpretation as a population-averaged model. The marginal distribution of Y_{ij} can be written as

$$Y_{ij} = \mu + \beta X_{ij} + \varepsilon_{ij}^*, \tag{2}$$

where $\text{cov}(\varepsilon^*) = \text{cov}(\mathbf{Y}) = \sigma_e^2 I + \sigma_A^2 J$, I is the $n = \Sigma n_i$ dimensional identity matrix, and J is the $n \times n$ block diagonal matrix with each element of the ith block equal to one. Thus, as pointed out by Ware (1985) and Zeger et al. (1988), β also measures the effect of the covariate X averaged over the units in the population and so has a direct interpretation in both cluster-specific and population-averaged terms. For other links, the distinction between these two models is more important.

In this paper, we compare cluster-specific and population-averaged approaches for binary data with respect to parameter interpretation. For this purpose, we examine one representative of each approach: the mixed-effects logistic model (Stiratelli et al. 1984; Anderson & Aitkin, 1985), a cluster-specific approach, and the GEE approach of Liang & Zeger (1986), a population-averaged approach. We describe the two approaches and compare them for the case of a single covariate X; the results generalize easily to the case of several covariates.

The mixed-effects logistic model is a generalization of the standard logistic model in which the intercept terms α_i are allowed to vary between clusters according to a distribution with density $f_\theta(\alpha)$. Within the ith cluster, the Y_{ij} are independent with

$$\text{logit } P(Y_{ij} = 1 \mid \alpha_i, X_{ij}) = \alpha_i + \beta X_{ij}. \tag{3}$$

Thus, β measures the change in the conditional logit of the probability of response with the covariate X for individuals in each of the underlying risk groups described by α_i.

The GEE approach on the other hand, specifies the marginal or population-averaged distribution of the Y_{ij}. Thus, it is assumed, for example, that

$$\text{logit } P(Y_{ij} = 1 \mid X_{ij}) = \alpha^* + \beta^* X_{ij} \tag{4}$$

together with some working covariance structure for $\mathbf{Y}$ to account for intraclass correlation. In this model, β^* measures the change in the logit of the proportion with $Y = 1$ for a unit increase in X. Liang & Zeger (1986) proposed estimation of β^* by solving a 'score-like' function which they call the generalized estimating equation (GEE). They show that the solution to the GEE, $\hat{\beta}_{\text{GEE}}$, is consistent and asymptotically normal, and give

an estimate of $\text{Cov}(\hat{\beta}_{\text{GEE}})$ which is consistent even when the assumed or working covariance matrix, $\text{Cov}(\mathbf{Y})$, is misspecified.

Although the marginal model (4) does not specify a unique mixed effects model, the mixed effects model (3) does specify a marginal model for Y_{ij},

$$P(Y_{ij}=1 \mid X_{ij}) = \int (1+e^{-\alpha-\beta X_{ij}})^{-1} f_\theta(\alpha)\, d\alpha, \tag{5}$$

which is not of the binary logistic form. Thus, although (3) and (4) appear similar, they are incompatible models and, unlike the linear models discussed above, their parameters have different interpretations. We make comparisons algebraically in § 2, geometrically in § 3, and in an example in § 5. In § 4, we discuss the relationship of this problem to that of omitted covariates in binary regression.

Zeger et al. (1988) also investigated cluster-specific and population-averaged models for binary data. They presented an approximation that is qualitatively similar to ours although it is derived under the assumption that the random effects distribution is normal. In this paper we present proofs and approximations which are valid for any random effects distribution; in addition, we attempt to develop intuition about the differences in parameter meaning for the two models and about the magnitude of the differences.

2 Model Comparisons

The population-averaged effect in the log odds scale from a unit increase in the covariate X is, by definition,

$$\beta_{\text{PA}}(X) = \log \frac{P(Y=1 \mid X+1)/P(Y=0 \mid X+1)}{P(Y=1 \mid X)/P(Y=0 \mid X)}. \tag{6}$$

For the population-averaged model (4), $\beta_{\text{PA}}(X)=\beta^*$, independent of X. For the cluster-specific approach, however, $\beta_{\text{PA}}(X)$ depends on X and, from (5) and (6), is

$$\beta_{\text{PA}}(X) = \log\left(\frac{E\{(1+e^{-\alpha-\beta(x+1)})^{-1}\}E\{(1+e^{\alpha+\beta x})^{-1}\}}{E\{(1+e^{\alpha+\beta(x+1)})^{-1}\}E\{(1+e^{-\alpha-\beta x})^{-1}\}}\right), \tag{7}$$

where the expectations are with respect to the distribution $f_\theta(\alpha)$ of α.

For random-effects distributions of interest, the integration necessary to evaluate (7) is intractable. However, the right-hand side of (7) can be approximated by expanding in a Taylor series about $\beta=0$. After some simplification, it can be seen that for β near 0, (7) is approximately independent of X with

$$\beta_{\text{PA}}(X) \doteq \beta\left\{1-\frac{\text{Var}(p)}{E(p)E(q)}\right\} = \beta[1-\rho(0)], \tag{8}$$

where $\text{logit}(p)=\alpha$, $q=1-p$, and $\rho(0)=\text{Corr}(Y_{ij}, Y_{ij'} \mid \beta=0)$ under the mixed-effects model. Thus $\rho(0)$ is the intracluster correlation among the $\mathbf{Y}$, when the covariate has no effect. It should be noted that, although the linear approximation does not depend on X, the quadratic approximation and $\beta_{\text{PA}}(X)$ do.

Equation (8) can also be derived from an information theoretic viewpoint. The results of White (1982) show that if the cluster-specific model (5) holds, then consistent estimates of the parameters of the marginal model (4) will converge to values α^* and β^* which minimize the Kullback–Leibler divergence (Kullback, 1959) between models (4) and (5). That is the values α^* and β^* minimize

$$E[\log\{P_{\text{CS}}(Y \mid X)/P_{\text{PA}}(Y \mid X)\}],$$

where the expectation is taken with respect to the cluster-specific model (5) while P_{CS} and P_{PA} refer to the probability distributions under models (5) and (4) respectively. This approach results in the following relationship between the parameters of (4) and those of (5):

$$(1+e^{-\alpha^*-\beta^*X})^{-1} = E(1+e^{-\alpha-\beta X})^{-1}, \tag{9}$$

where the expectation is taken with respect to the distribution of α as in (5). Expanding the logit of (9) in a Taylor series about $\beta = 0$ yields

$$\alpha^* + \beta^* X \doteq \log\left[\frac{E(p)}{E(q)}\right] + \beta X\left[1 - \frac{\mathrm{Var}\,(p)}{E(p)E(q)}\right], \tag{10}$$

where p and q are given above. Thus, the marginal distribution induced by the mixture model (5) is approximately logistic with an attenuated slope coefficient given by (8). Equation (8) has also been derived by Gail (1988, eqn (13)), in a different context.

Since $0 \leq \rho(0) \leq 1$, (8) suggests that, at least for small β, $|\beta_{PA}(X)| < |\beta|$, so that the population-averaged effect is smaller than the cluster-specific effect. (In the next section we show that this suggestion is true generally.) It is also clear from (8) that large values of $\mathrm{Var}\,(p)$ (or equivalently $\mathrm{Var}\,(\alpha)$) lead to large differences between β and $\beta_{PA}(X)$. It is easily seen that $\beta = 0$ implies $\beta_{PA} = 0$. Thus, if the cluster-specific model is true, tests of $\beta = 0$ using a population-averaged approach, and variance estimates appropriate for the cluster design, will have the correct Type I error level but will tend to be less efficient than those using a cluster-specific approach. Finally, if there are no random effects, so that $\mathrm{Var}\,(p) = 0$, then $\beta = \beta_{PA}(X)$ and the model (3) also specifies (4) for the marginal mean of Y_{ij}.

Zeger et al. (1988) also present an attenuation formula analogous to (8) for the special case in which the random effects α follows a normal distribution. Their results are qualitatively similar to ours for this special case. Their expression also suggests that the parameters of the population-averaged approach will be closer to zero than those of the cluster-specific approach, but it lacks the intuitive appeal of (8).

3 Geometric Approach

3.1 Random Intercepts

Consider the cluster-specific model (3); let α be a particular value of the random effect, and let $\psi = \exp(\beta)$ be the cluster-specific odds ratio. If $P_l(\alpha) = \Pr\{Y = 1 \mid X + l, \alpha\}$ for $l = 0, 1$, it follows that

$$\Delta(\alpha) = P_1(\alpha) - P_0(\alpha) = \frac{(\psi - 1)P_0(\alpha)[1 - P_0(\alpha)]}{1 - P_0(\alpha) + \psi P_0(\alpha)}, \tag{11}$$

where throughout X is considered as fixed and dependence on ψ and X is suppressed. Without loss of generality, we consider $\psi > 1$ and note that for each fixed ψ, (11) defines a curve in the (P_0, Δ) plane with $\Delta \geq 0$ for $0 < P_0 < 1$ and $\partial^2\Delta/\partial P_0^2 < 0$ for $0 < P_0 < 1$. Further, the concave regions defined by these curves are nested as ψ increases from 1 to ∞; that is, $\partial\Delta/\partial\psi > 0$ for $\psi > 1$. Figure 1 provides an illustration.

Again with $\psi > 1$ and fixed, the corresponding population-averaged model is obtained by averaging with respect to the distribution $f_\theta(\alpha)$ of α. This gives rise to the point (P_0^{PA}, Δ^{PA}) in the concave region defined by ψ where $P_0^{PA} = E(P_0(\alpha))$ and $\Delta^{PA} =$

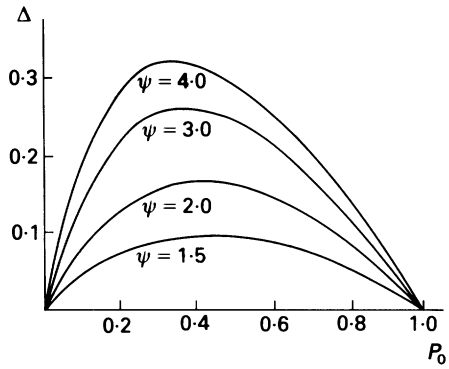

Figure 1. Plots of Δ versus P_0 for given odds ratio ψ from equation (9).

$E[P_1(\alpha) - P_0(\alpha)]$. The corresponding population-averaged odds ratio then satisfies

$$\psi_{PA} = \exp(\beta_{PA}) = \frac{P_1^{PA}/[1 - P_1^{PA}]}{P_0^{PA}/[1 - P_0^{PA}]},$$

where $P_1^{PA} = P_0^{PA} + \Delta^{PA}$. It follows that ψ_{PA} is the value of ψ that defines the curve through (P_0^{PA}, Δ^{PA}). Since $\partial^2\Delta/\partial P_0^2 < 0$, we have $E[\Delta(P_0)] < \Delta[E(P_0)]$, by Jensen's inequality. Since Δ is an increasing function of ψ, it follows immediately that $1 < \psi_{PA} < \psi$ and further that $\psi_{PA} = \psi$ if and only if var $(\alpha) =$ var $(P_0) = 0$.

This establishes the following theorem.

THEOREM 1. *In the cluster-specific logistic model* (3) *with mixing density* f_θ *with* var $(\alpha) > 0$, $|\beta_{PA}(X)| < |\beta|$ *with equality if and only if* $\beta = 0$.

It is intuitively clear from this geometric argument that the greater the variation in the distribution of α, the larger the difference between $|\beta_{PA}|$ and $|\beta|$ is expected to be. In the extreme case, β_{PA} can be made arbitrarily close to zero by assigning all the mass in the distribution of α to values of P_0 very near 0 and very near 1 (α near $-\infty$ and ∞). It can be seen from Fig. 1 that there will be little difference between β and β^{PA} provided the distribution of P_0 is fairly concentrated (has small variance) or if only values of P_0 that are small (near 0) can occur with appreciable probability. In this latter case, the logistic model is a good approximation to the relative risk model where, as discussed below, the parameters have both a population-averaged and a cluster-specific interpretation. Similar remarks hold if P_0 must be near 1.

The approach of this section can also be used to show that the parameters of models for the relative risk, $\omega = P_1(\alpha)/P_0(\alpha)$, have both a cluster-specific and population-averaged interpretation. Solving for P_1 in terms of the relative risk and P_0 we have $P_1(\alpha) = \omega P_0(\alpha)$, so that

$$\Delta'(\alpha) = P_1(\alpha) - P_0(\alpha) = (\omega - 1)P_0(\alpha).$$

Let $\Delta'^{PA} = E[\Delta'(\alpha)]$. Since Δ' is a linear function of P_0, the point $(P_0'^{PA}, \Delta'^{PA})$ is on the line determined by the cluster-specific relative risk ω. Thus, the population-averaged relative risk is $\omega_{PA} = P_1^{PA}/P_0^{PA} = \omega$, the cluster-specific relative risk.

3.2 Random Slopes

Equation (10) extends easily to the case where the regression parameters are random. Within the ith cluster, the Y_{ij} are taken to be independent with

$$\text{logit } P(Y_{ij} \mid \alpha_i, b_i, X_{ij}) = \alpha_i + (\beta + b_i)X_{ij}, \tag{12}$$

where α_i and b_i vary between clusters according to a distribution with multivariate density $f_\theta(\alpha, b)$ and $E(b) = 0$.

Minimizing the Kullback–Leibler divergence between models (4) and (12) at a fixed X, as in § 2, leads to the following expression relating the parameters of the marginal model (4) to those of the random-slopes model (12):

$$(1 + e^{-\alpha^* - \beta^* X})^{-1} = E(1 + e^{-\alpha - (\beta + b)X})^{-1}, \tag{13}$$

where the expectation is taken with respect to the joint distribution of α and b. Expanding the logit of the right-hand side of (13) in a Taylor series about $\beta = 0$ yields

$$\alpha^* + \beta^* X \doteq \log\left[\frac{Ep(X)}{Eq(X)}\right] + \beta X\left[1 - \frac{\operatorname{Var} p(X)}{Ep(X)Eq(X)}\right], \tag{14}$$

where logit $p(X) = \alpha + bX$, $q(X) = 1 - p(X)$ and expectations are with respect to the joint distribution of α and b.

Unlike the fixed slope situation, the coefficients of both the 'intercept' and βX depend on X. This is similar to an expression given by Zeger et al. (1988) who also present an approximation of this type. Neither their approximation for (α, b) bivariate normal nor (14) is particularly helpful however since they do not describe the overall relationship between the parameters of the two models.

In order to get additional insight, it is possible to consider a distribution for the covariate X. In this case, the Kullback–Leibler divergence can be simultaneously minimized in α^* and β^*. This leads to the two equations

$$E[X^j\{1 + e^{-\alpha^* - \beta^* X}\}^{-1}] = E[X^j\{1 + e^{-\alpha - (\beta + b)X}\}^{-1}] \quad (j = 0, 1), \tag{15}$$

where the expectation is taken with respect to the joint distribution of α, b and X. Solutions in α^* and β^* to these equations could be examined in particular cases, but simple general statements regarding the relationships between β^* and β do not seem to be available. Further progress can be made in the two sample case with random slopes, for example as discussed below. In the fixed slope case, when α is independent of X, the solution of (15) in β^* satisfies to first order the attenuation formula (8).

To generalize the geometric interpretation to the random slopes model we would consider the following generalization of (11)

$$\Delta(\alpha, b) = P_1(\alpha, b) - P_0(\alpha, b) = \frac{(\psi e^b - 1)P_0(1 - P_0)}{1 - P_0 + \psi e^b P_0},$$

where $P_l(\alpha, b) = P(Y = 1 \mid X + l, \alpha, b)$. Without loss of generality, we consider $\psi > 1$. However, for the random slopes model, the geometric proof of Theorem 1 does not extend in general due to the nonconcave nature of $\Delta(\alpha, b)$.

Theorem 1 would hold if $E[\Delta(P_0, b)] \leq \Delta[E(P_0), E(b)]$, where expectations are with respect to the joint distribution of P_0 and b. For $b > -\log \psi$, $\partial^2\Delta/\partial P_0^2 < 0$, so that

$$E[\Delta(P_0, b)] = E_b E_{P_0|b}[\Delta(P_0, b)] \leq E_b \Delta[E(P_0 \mid b), b].$$

Theorem 1 would hold for $b > -\log \psi$ if $\Delta[E(P_0 \mid b), b]$ were a concave function of b (or e^b). Since P_0 and b are correlated, the concavity of $\Delta[E(P_0 \mid b), b]$ cannot be established for an arbitrary multivariate mixing density $f_\theta(\alpha, b)$. For a specified f, one can calculate $E(P_0 \mid b)$ and examine the concavity of Δ.

Theorem 1 holds for the important special case in which α and b are uncorrelated and the covariate X can only assume the values 0 and 1. For this case, P_0 and b are uncorrelated and it is easy to show that $\Delta[E(P_0), b]$ is concave in e^b.

In many applications, the restriction $b > -\log \psi$ is desirable. If, for example, $b < -\log \psi$ is possible, this would suggest that, although the overall association between the covariate and the outcome is positive ($\psi > 1$ or $\beta > 0$), there exists a subgroup in the population for which the covariate outcome association is negative. In models with random slope, we typically wish to allow variation in the degree of association, but expect the association to be always in one direction.

4 Analogy with Omitted Covariates

In the above, the random effects, α, can be thought of as an omitted covariate, and there is a close connection between the results above and those relating to omitted covariates in binary regression models. Lee (1982) and Gail, Wieand & Piantadosi (1984) have shown that estimates of the effect of a randomly assigned treatment will be biased toward no association unless the included and omitted covariates are uncorrelated, conditional on the response Y. Thus, from these papers, we would expect that population-averaged effects would be closer to zero than cluster-specific effects. In fact, under the mixed effects model (3), it is easy to show that the conditional density of α, given the response Y and the covariate X depends on X unless var $(\alpha) = 0$, that is unless there are no random effects.

Gail et al. (1984) also gave an expression (equation (2.9)) for the magnitude of the bias in estimation of the effect of the included covariate due to an omitted covariate. Rather than expand about $\beta = 0$, as we do in (7), they considered a Taylor expansion about $\gamma = 0$, where γ is the effect of the omitted covariate. (This is analogous to an expansion of (6) about var $(\alpha) = 0$.) The qualitative findings of (8) and those of their equation (2.9) are the same. For example, both approximations suggest that the population-averaged effect of X (the effect of X with an omitted covariate) will be closer to no association than the cluster-specific effect of X (the true underlying effect of X). In addition, they show that the differences in magnitudes of the effects increase with var (α).

Gail et al. (1984) and B.V. Bye and J.M. Dykacz (unpublished 1987 presentation) report the results of simulations which assess the bias due to omitted covariates in logistic regression. They generate data with two independent covariates with unit variances, and fit models with one. Table 1 presents the values of the parameters of the true underlying logistic models, the biases obtained from the simulations of Bye and Dykacz and the biases predicted by (8) and equation (2.9) of Gail et al. (1984) and the approximation of Zeger et al. (1988). For γ near zero, the biases predicted by all three correspond closely to the simulated biases. For $\gamma = -4{\cdot}0$, the predicted biases of (8) and of Zeger et al. are close to the simulated bias, whereas the bias predicted by equation (2.9) is not.

Table 1

Simulated biases (as reported by Bye and Dykacz) and predicted biases in estimation of β for three methods

μ	β	γ	Simulated bias in β	Predicted bias from (8)	Predicted bias Gail et al. (1984)	Predicted bias Zeger et al. (1988)
1·0	1·0	0·5	−0·068	−0·056	−0·048	−0·041
0·5	0·3	0·2	0·010	−0·003	−0·003	−0·002
1·0	4·0	−4·0	−2·440	−2·540	−7·570	−2·440

5 An Example

In this section, cluster-specific and population-averaged models are fitted to data from an ongoing study of breast disease conducted at the University of California, San Francisco. One component of the study consists of obtaining a sample of fluid from both breasts of all study women. Ernster et al. (1987) give details of study design and techniques.

In the analyses given here, the binary outcome was whether a sample of breast fluid could ($Y = 1$) or could not ($Y = 0$) be obtained from each breast and the covariates considered were $X_2 =$ age in years, $X_3 =$ age at menarche in years, a binary indicator $X_4 = 1$ if the woman was parous and $X_4 = 0$ if not, and a binary indicator of whether ($X_1 = 1$) or not ($X_1 = 0$) physical examinations of each breast found evidence of dysplasia. The sample comprised 490 white, premenopausal women who had no breast disease.

We fit the model (3) with the random effects assumed to follow a 21-point binomial approximation to the normal distribution. We used the software package EGRET which maximizes the likelihood using a quasi-Newton algorithm; standard errors are estimated from the sample information matrix. We also fit a population-averaged logistic model accounting for correlation between breasts using GEE (Liang & Zeger, 1986) with the exchangeable correlation structure. Standard errors were calculated using the robust variance estimators.

Table 2 shows that, as predicted by Theorem 1 and (8), the population-averaged regression coefficients are closer to zero than the cluster-specific coefficients. The results of the population-averaged analysis suggest that the logit of prevalence of fluid availability is 0·435 units larger among women who were parous than among women who were nulliparous. The other regression coefficients of the population-averaged approach have similar interpretations. On the other hand, the results of the cluster-specific analysis suggest that for women in the same latent risk group (same α_i) the logit of the probability of fluid availability is 1·25 units higher for those women who are parous.

Table 2 also gives the ratio of the estimated population-averaged coefficients to the estimated cluster-specific coefficients. The Taylor expansion (8) predicts this ratio to be $1 - \mathrm{Var}\,(p)/[E(p)E(q)]$, which can be estimated from the fitted distribution of α. The estimate of 0·35 is nearly identical to the observed ratios in Table 2; for these data, the Taylor approximation (8) is very accurate. The approximation of Zeger et al. (1988) yields a similar estimate of 0·37.

It is of some interest to note that the standardized coefficients ($\hat{\beta}/\mathrm{SE}(\hat{\beta})$) are nearly the same for the two models. Thus, for example, Wald type tests for covariate effects would generate very similar significance levels for the populations-averaged and for the cluster-specific models.

Table 2

Point estimates (standard errors) of regression coefficients for cluster-specific and population-averaged approaches applied to data on the availability of breast fluid.

Variable	Mixed effects, β_{CS}	Liang–Zeger, β_{PA}	Ratio, β_{PA}/β_{CS}
Intercept	−1·951 (2·49)	−0·620 (0·908)	
Dysplasia	0·609 (0·462)	0·208 (0·155)	0·351
Age	0·146 (0·044)	0·050 (0·015)	0·343
Age at menarche	−0·414 (0·172)	−0·147 (0·059)	0·355
Full term birth	1·249 (0·610)	0·435 (0·216)	0·348
std. dev. (α_i)	4·218 (0·511)		

6 Discussion

6.1 Interpretation of Coefficients

Although the cluster-specific model seems to provide the more unified approach, parameter interpretation in these models is difficult. Consider for example, the coefficients for parity in § 5. The cluster-specific model presupposes the existence of latent risk groups indexed by α_i, and parameter interpretation is with reference to these groups. No empirical verification of this statement can be available from the data unless the latent risk groups can be identified. Since each individual is assumed to have her own latent risk α_i, the model almost invites an unjustified causal statement about the change in the odds of fluid availability for a given woman who ceases to be nulliparous.

With covariates that vary within clusters, such as dysplasia, the mixed effect model provides a more satisfactory interpretation. The estimated coefficient for dysplasia, however, involves both a within cluster comparison based on differences between breasts for the same woman, and a between cluster comparison of average levels. An alternative and perhaps preferable analysis would partition dysplasia into average level of dysplasia for between cluster comparisons and departures from average for within cluster comparisons and allow separate coefficients for these two covariates. Population-averaged comparisons, on the other hand, make no specific use of within cluster comparisons for cluster varying covariates and substantially underestimate within cluster risks. Related to this, population-averaged models cannot provide estimates of changes within individuals over time; these are often quantities of central interest in longitudinal studies.

6.2 Estimation of the Mixing Distribution

Both models (3) and (4) allow for extra-binomial variation in the cluster totals. In (4), the extra-binomial variation is allowed in the specification of the covariance matrix. In (3), the random effects, α_i, account for extra-binomial variation, but as the variability of α_i varies, both the mean and the variance of the cluster totals are affected. As a consequence, estimation of the dispersion of the mixing distribution is affected by departures of the empirical dose response of cluster averages from the logistic form as well as by the departure from binomial variation of the variance of these responses for given covariate values. In the mixed effects model, lack of fit is confounded with extra-binomial variation. Estimation of the mixing distribution is highly dependent upon the assumed logistic link in (3). This a serious difficulty with the model (3); its specification involves the introduction of non-testable assumptions and suitable care is required in the interpretation of the results obtained. For covariates measured on the cluster, it is useful to adjust estimates to describe population-averaged dose response relationships since these are empirically verifiable. The result (8) provides a first step in this direction.

The analogy to omitted covariates provides one method of interpreting cluster-specific and population-averaged models. We can consider the random effect in a cluster-specific model to represent the totality of cluster-constant covariate effects omitted from the model that are orthogonal to those already present. As these additional orthogonal covariates are added to the model, the cluster-specific model will tend toward one in which $\text{var}(\alpha) = 0$ and so would admit a population-averaged interpretation for the regression variables. In the population-averaged model, the effects are based on averaging over subpopulations defined by the specific covariates in the model. As one

adds covariates to the model, the attenuation due to omitted covariates (or equivalently, due to the population-averaging) will be reduced; the averaging is done over ever smaller groups with a particular risk group and the corresponding cluster-specific measure of effect as the limit.

Cluster-specific models are best equipped to address questions relating to modification of a particular cluster (and hence must be used with caution unless the variable of interest is cluster varying and experimentally manipulated) since population-averaged models do not estimate this effect. In much of what has been done here, it is implicitly assumed that the mixed-effects model (3) is appropriate and implications for population-averaged results are investigated. The obverse problem, in which the true model is of the population-averaged type (e.g. beta-binomial) and a mixed-effects model is used for analysis, could give additional insights.

6.3 Odds Ratios and Interpretation

The problems with interpretation arise as a consequence of the use of odds ratios to measure excess risk. Although these have become standard measures in epidemiologic work, it needs to be borne in mind that the odds ratio itself is insufficient information upon which to base decision making or to assess seriousness of the risk. Thus an odds ratio of 2 could quite rationally be ignored if the baseline risk were one in a million, but must be taken very seriously if the baseline risk is one in four. The odds ratios from the cluster-specific and population-averaged models, although quite different in magnitude, are in fact describing exactly the same dependence of risk on the covariates. Neither can truly be interpreted or understood without reference to the baseline risk. In the population-averaged case, the baseline risk is simply the proportion of positive responses at the standard level of the covariate; in the cluster specific case, the baseline risk is in fact a distribution of risks. If the separate risk groups can be identified, then cluster-specific coefficients have meaning to individuals who can identify their baseline risk and the associated odds ratio. But if the risk groups are not identifiable, the average baseline risk together with the corresponding average odds ratio is an appropriate brief summary. With this convention, the two analyses provide the same qualitative assessment.

Acknowledgements

The authors would like to thank Dr. Nicholas Petrakis of the Department of Epidemiology and Biostatistics, University of California, San Francisco for providing the data set used in the example. This research was supported by a grant from the National Institute on Drug Abuse co-ordinated by the Societal Institute of the Mathematical Sciences.

References

Altham, P.M.E. (1978). Two generalizations of the binomial distribution. *Appl. Statist.* **27,** 162–167.

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

Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research,* **1**: *The Analysis of Case-Control Studies.* Lyon: International Agency for Research on Cancer.

Ernster, V.L., Wrensch, M.R., Petrakis, N.L., King, E.B., Miike, R., Murai, J. & Siiteri, P.K. (1987). Benign and malignant breast disease: Initial study results of serum and breast fluid analyses of endogenous estrogens. *J. Nat. Canc. Inst.* **79,** 949–960.

Gail, M.H. (1988). The effect of pooling across strata imperfectly balanced studies. *Biometrics* **44,** 151–162.

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

Kullback, S. (1959). *Information Theory and Statistics.* New York: Wiley.

Kupper, L.L., Portier, C., Hogan, M.D. & Yamamoto, E. (1986). The impact of litter effects on dose-response modeling in teratology. *Biometrics* **42,** 85–98.
Lee, L.F. (1982). Specification error in multinomial logit models: Analysis of the omitted variable bias. *J. Econometrics* **20,** 197–209.
Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73,** 13–22.
Prentice, R.L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Am. Statist. Assoc.* **81,** 321–327.
Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44,** 1033–1048.
Stiratelli, R., Laird, N. & Ware, J.H. (1984). Random-effects model for serial observations with binary response. *Biometrics* **40,** 961–971.
Ware, J.H. (1985). Linear models for the analysis of serial measurements in longitudinal studies. *Am. Statistician* **39,** 95–101.
Ware, J.H., Lipsitz, S. & Speizer, F.E. (1988). Issues in the analysis of repeated categorical outcomes. *Statist. Med.* **7,** 95–108.
White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50,** 1–25.
Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31,** 949–952.
Zeger, S.L., Liang, K.-Y. & Albert, P.A. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44,** 1049–1060.

Résumé

Des échantillons corrélés ou en grappes de réponses binaire se présentent fréquemment en pratique à cause de mesures répétées ou de sous-échantillonage d'échantillons initiaux. La corrélation intra-grappe dans les problèmes de régression binaires a été l'objet d'études et méthodes récentes. On peut citer les méthodes spécifiques aux grappes telles que celles basées sur les modèles logistiques à effets mixtes ainsi que les méthodes de moyennes de populations telles que celles basées sur les modèles bêta-binomiaux. Notre article considère l'interprétation des paramètres de régression dans ces méthods générales. Nous démontrons que, contrairement à ce qui se passe pour des résultats Gaussiens corrélés, pour des données binaires corrélées les paramètres spécifique aux grappes et ceux des modèles de moyennes de populations décrivent différents types d'effet des covariables sur les probabilités de réponse. Dans le cas des valeurs à zéro aléatories, nous démontrons que si la modèle des grappes est verifié, mesurer l'effet des covariables par la méthode des moyennes de populations donne des résultats plus proches de zéro que si l'on mesure par les méthodes spécifiques aux grappes. On démontre aussi que la différence entre les résultats des deux méthodes augmente avec la corrélation intra-grappe. Nous étudions aussi le cas des pentes aléatories. Ces résultats s'appliquent à toute distribution des effets aléatoires. On le décrit pour des données sur la capacité d'obtenir des échantillons de liquide mammaire.

[*Received July* 1989, *accepted June* 1990]