BIOMETRICS 40, 961–971
December 1984

Random-Effects Models for Serial Observations with Binary Response

Robert Stiratelli

Computer Applications Research Department, Rohm and Haas Research Laboratories, 727 Norristown Road, Spring House, Pennsylvania 19477, U.S.A.

and

Nan Laird and James H. Ware

Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

SUMMARY

This paper presents a general mixed model for the analysis of serial dichotomous responses provided by a panel of study participants. Each subject's serial responses are assumed to arise from a logistic model, but with regression coefficients that vary between subjects. The logistic regression parameters are assumed to be normally distributed in the population. Inference is based upon maximum likelihood estimation of fixed effects and variance components, and empirical Bayes estimation of random effects. Exact solutions are analytically and computationally infeasible, but an approximation based on the mode of the posterior distribution of the random parameters is proposed, and is implemented by means of the EM algorithm. This approximate method is compared with a simpler two-step method proposed by Korn and Whittemore (1979, *Biometrics* **35**, 795–804), using data from a panel study of asthmatics originally described in that paper. One advantage of the estimation strategy described here is the ability to use all of the data, including that from subjects with insufficient data to permit fitting of a separate logistic regression model, as required by the Korn and Whittemore method. However, the new method is computationally intensive.

1. Introduction

Longitudinal (or panel) studies which require repeated observations on participants are an integral part of medical and biological science. They have been fundamental to the study of physical and cognitive development (Goldstein, 1979; Nesselroade and Baltes, 1979) and normative aging. They are important in the study of chronic diseases such as arthritis, nephritis, diabetes and chronic obstructive pulmonary diseases. Panel studies are also used increasingly in assessing the health effects of air pollution (Stebbings and Hayes, 1976).

Statistical methodology for these studies has been developed primarily for the analysis of repeated measurements. Multivariate and repeated-measures analysis of variance (Bock, 1975; Grizzle and Allen, 1969) and growth-curve analysis (Rao, 1975; Ware and Wu, 1981) have been used successfully with balanced designs. Laird and Ware (1982) and Strenio, Weisberg and Bryk (1983) recommend empirical Bayes analysis based on random-effects models. The latter methods are easily implemented with unbalanced designs or incomplete data.

Methods for the analysis of longitudinal data with binary, or more generally, polychotomous response, are far less well developed. Methods which have been proposed perform poorly in settings with many covariates, unbalanced data, missing observations

Key words: Empirical Bayes; Binary response; EM algorithm; Longitudinal data; Random logistic model.

and/or large numbers of observations per subject (Markus, 1979; Koch *et al.*, 1977; Hasselblad, 1978). For serial dichotomous response, Korn and Whittemore (1979) proposed a logistic growth-curve model with normally distributed random coefficients which has many of the strengths of the corresponding random-effects models for measured response. A drawback of the logistic-normal model is its analytic intractability. Korn and Whittemore proposed an approximate analysis based on a separate logistic regression for each subject. Their method requires the exclusion of individuals with low response rates, missing data, or small numbers of repeated observations.

This paper generalizes the logistic-normal model for dichotomous response and develops an empirical Bayes approach to inference. The benefits of this approach are described, as well as difficulties in its implementation. After presenting the model and method of analysis, we apply it to a set of panel data arising from a study of the impact of air pollution on asthma attack rates. These data were also analyzed by Korn and Whittemore (1979) and we compare the results for the two methods of analysis. The two methods lead to somewhat different results regarding the effect of air pollution on asthma attack rates, apparently because the model simplification and subject selection required by the Korn–Whittemore method lead to residual confounding and selection bias.

2. The Model

The data in a general longitudinal study can be characterized as follows. Let $\mathbf{y}_i$ denote an $n_i \times 1$ vector of responses for the ith individual, $i = 1, \ldots, N$. Each element, y_{ij}, is assumed to be a binary response; extension to categorical responses is straightforward. Let $p_{ij} = \text{pr}(y_{ij} = 1)$, $\lambda_{ij} = \text{logit}\, p_{ij}$, and let $\boldsymbol{\lambda}_i$ denote the $n_i \times 1$ vector of logits. Each individual also has a set of covariates, some of which vary over occasions of measurement (e.g. air pollution, weather variables, day of week, etc.) and others which are fixed (sex, initial conditions, etc.). Korn and Whittemore (1979) proposed a version of the following two-stage model based on logistic regression. Let $\mathbf{Z}_i$ denote an $n_i \times k$ matrix of the occasion-varying (or within-person) covariates plus an intercept. Assume $\boldsymbol{\lambda}_i = \mathbf{Z}_i\boldsymbol{\nu}_i$, where $\boldsymbol{\nu}_i$ is an unknown parameter vector for the ith individual. Conditional on $\boldsymbol{\nu}_i$, assume that the likelihood for the ith individual is

$$l_i = \prod_{i=1}^{n_i} p_{ij}^{y_{ij}}(1 - p_{ij})^{1-y_{ij}}. \tag{2.1}$$

At Stage 2 of the model, each $\boldsymbol{\nu}_i$ is taken to be $\text{MVN}(\mathbf{W}_i\boldsymbol{\alpha}, \mathbf{D})$, where $\mathbf{W}_i$ is a $k \times p$ matrix of between-subject covariates, that is, covariates that vary over subjects but remain constant over occasions for a given subject.

Several features of this model deserve special attention. Stage 1 allows modeling of the within-subject variation (or occasion-to-occasion variation) separately for each subject, as a function of the within-person covariates. Thus each individual has a unique response function whose parameters are specified by $\boldsymbol{\nu}_i$. At Stage 2 we model the between-subject variation by postulating a distribution for the individual growth-curve parameters. The mean values of the distribution of random effects can depend upon the between-subject covariates.

Observations on different subjects are assumed to be independent. Serial correlation between successive measurements on the same subject can arise in one of two ways. First, serial correlation which arises as the result of true state dependence can be modeled at Stage 1, by allowing $\mathbf{Z}_i$ to include a covariate which is the value of the response on the previous occasion. For example, a modulated two-state Markov chain is specified for each individual at Stage 1 by taking

$$\lambda_{ij} = \nu_{1i} + \nu_{2i}y_{i,j-1} + \sum_{m=3}^{k} \nu_{mi}Z_{mji}.$$

Here p_{ij} is the *conditional* probability of response on Occasion j, given the covariates $\mathbf{Z}_i$, parameter vector $\boldsymbol{\nu}_i$, and previous responses, $y_{i,j-1}$, $y_{i,j-2}$, etc. The Markov assumption implies that the probability of response depends on past history only through the response on the previous occasion. The response probability is modulated by the effects of other occasion-varying covariates, Z_{3ji}, Z_{4ji}, etc.

Serial correlation which is the result of heterogeneity in the population parameters (as opposed to state dependence) is introduced at Stage 2. Even if we assume independence for successive observations on an individual at Stage 1, correlations between observations on the same individual will still be induced at Stage 2 because parameters vary over individuals. Such 'conditional independence' models have certain advantages over the Markov ones. Technically, missing data, nonequispaced intervals and specification of initial conditions can cause problems with Markov models. More importantly, the interpretation of covariate effects becomes more difficult in Markov models. The exponentiated parameters for covariate effects still may be interpreted as relative odds (in the right scale) but they are conditional on previous responses. If there is considerable true state dependence, the conditional effects of other covariates may be small even though they may have considerable impact on the marginal response rates. By using a conditional-independence model, we can study the effects of covariates on the marginal response rates (conditional only on $\mathbf{Z}_i$). In many settings this may be more desirable. It is also analogous to growth-curve analysis in the measured-response setting.

We refer to the preceding model as a growth-curve model, since a different regression model is fitted for each individual at Stage 1. This paper considers a larger and more flexible family of models referred to as the general logistic-linear mixed model. At Stage 1 let

$$\boldsymbol{\lambda}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i, \tag{2.2}$$

and at Stage 2 assume that $\mathbf{b}_i$ is MVN($\mathbf{0}$, $\mathbf{D}$). These assumptions define the general mixed linear model for the logits of the response probabilities. Taking $\mathbf{X}_i = \mathbf{Z}_i\mathbf{W}_i$ and $\mathbf{b}_i = \nu_i - \mathbf{W}_i\alpha$ yields the original growth-curve model as a special case. To implement the empirical Bayes approach to inference for Model (2.2), we will further assume that $\boldsymbol{\alpha}$ has a diffuse prior distribution. This is accomplished by assuming at Stage 2 that $\boldsymbol{\alpha}$ is also MVN($\mathbf{0}$, $\boldsymbol{\Gamma}$), and subsequently letting $\boldsymbol{\Gamma}^{-1} \rightarrow \mathbf{0}$.

3. Parameter Estimation for the General Logistic-Linear Mixed Model

In principle, one could estimate $\boldsymbol{\alpha}$ and $\mathbf{D}$ by maximizing the marginal likelihood of the data,

$$L(\boldsymbol{\alpha}, \mathbf{D}) = \prod_{i=1}^{N} \int_{\mathbb{R}^k} l_i \exp(-\tfrac{1}{2}\mathbf{b}_i^T\mathbf{D}^{-1}\mathbf{b}_i)|\mathbf{D}|^{-1/2}\, d\mathbf{b}_i, \tag{3.1}$$

where l_i is given in (2.1) and $\mathbf{p}_i$ is the vector of logits given in (2.2). Estimates of the random individual parameters would be obtained by using an empirical Bayes strategy:

$$\hat{\mathbf{b}}_i = \mathrm{E}(\mathbf{b}_i|\mathbf{y}_i, \hat{\boldsymbol{\alpha}}, \hat{\mathbf{D}}). \tag{3.2}$$

This approach has two drawbacks. First, maximum likelihood estimates of variance components are biased in small samples, the bias increasing as the dimension of $\boldsymbol{\alpha}$ increases relative to N. [See Harville (1977, p. 325) and references therein.] A standard fix-up for this problem in the measured-response case is the use of restricted maximum likelihood (REML) estimates of the variance components. In the balanced ANOVA setting, the REML analysis is equivalent to maximizing the likelihood obtained by integrating out the fixed effects using a diffuse prior (Laird and Ware, 1982).

We will adopt that strategy here, estimating $\mathbf{D}$ by maximizing

$$L_R(\mathbf{D}, \mathbf{\Gamma}^{-1} = \mathbf{0}) = \lim_{\mathbf{\Gamma}^{-1} \to 0} \left\{ \int_{\mathbb{R}^p} L(\boldsymbol{\alpha}, \mathbf{D}) \exp(-\tfrac{1}{2}\boldsymbol{\alpha}^T \mathbf{\Gamma}^{-1} \boldsymbol{\alpha}) |\mathbf{\Gamma}|^{-1/2}\, d\boldsymbol{\alpha} \right\}. \tag{3.3}$$

The associated estimate of $\boldsymbol{\alpha}$ is

$$\hat{\boldsymbol{\alpha}} = \mathrm{E}(\boldsymbol{\alpha} | \mathbf{y}, \hat{\mathbf{D}}, \mathbf{\Gamma}^{-1} = \mathbf{0}), \tag{3.4}$$

where $\mathbf{y}^T = (\mathbf{y}_1^T, \ldots, \mathbf{y}_N^T)$, and expectation is over the joint posterior distribution of $\boldsymbol{\alpha}$ and $\mathbf{b}^T = (\mathbf{b}_1^T, \ldots, \mathbf{b}_N^T)$. Variances for $\hat{\boldsymbol{\alpha}}$ and $\hat{\mathbf{b}}$ can be estimated by

$$\hat{\mathbf{\Sigma}}_\alpha = \mathbf{var}\,(\boldsymbol{\alpha} | \mathbf{y}, \hat{\mathbf{D}}, \mathbf{\Gamma}^{-1} = \mathbf{0}) \tag{3.5}$$

$$\hat{\mathbf{\Sigma}}_b = \mathbf{var}\,(\mathbf{b} | \mathbf{y}, \hat{\mathbf{D}}, \mathbf{\Gamma}^{-1} = \mathbf{0}). \tag{3.6}$$

Because of the equivalence of flat prior Bayes and maximum likelihood estimation, this general strategy for estimating $\boldsymbol{\alpha}$ and $\mathbf{D}$ could be regarded as maximum likelihood, with an REML adjustment for the variance components. The estimate of $\mathbf{b}_i$ is also a generalization of the standard method for handling estimates of random effects in the general linear mixed model. Laird and Ware (1982) discuss this empirical Bayes approach to inference in the case of measured response with linear models and Gaussian error structure. In that setting, there is a direct connection between empirical Bayes and standard frequentist strategies for estimation and inference. Optimality of the empirical Bayes strategy can be demonstrated by applying standard sampling-theory optimality criteria.

The second problem we encounter with either ML or empirical Bayes approaches is that closed-form expressions for the integrals in (3.1) and (3.3) [and those implicit in (3.2) and (3.4)–(3.6)] do not exist. Thus, standard analytical methods cannot be used to calculate $\hat{\mathbf{D}}$, and, given $\hat{\mathbf{D}}$, numerical integration in Nk dimensions is required to obtain $\hat{\boldsymbol{\alpha}}$ and $\hat{\mathbf{\Sigma}}_\alpha$. The proposed solution for estimating $\boldsymbol{\alpha}$ and $\mathbf{b}$ is to use posterior modes rather than means in (3.2) and (3.4). We approximate the posterior variances by the inverse of the information matrices based on the corresponding posterior distributions and evaluated at the mode. This is equivalent to approximating the posterior distribution of $\boldsymbol{\alpha}$ and $\mathbf{b}$ by a multivariate normal distribution that has the same mode and curvature at the mode as the true posterior. Denoting the joint posterior distribution of $\boldsymbol{\alpha}$ and $\mathbf{b}$ as $p(\boldsymbol{\alpha}, \mathbf{b} | \mathbf{y}, \mathbf{D}, \mathbf{\Gamma})$, define $\hat{\boldsymbol{\alpha}}$, $\hat{\mathbf{b}}$ and $\hat{\Sigma}_{\alpha b}$ by

$$p(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{b}}) = \sup_{\boldsymbol{\alpha}, \mathbf{b}} p(\boldsymbol{\alpha}, \mathbf{b} \mid \mathbf{y}, \hat{\mathbf{D}}, \mathbf{\Gamma}^{-1} = \mathbf{0}) \tag{3.7}$$

and

$$\hat{\mathbf{\Sigma}}_{\alpha b} = -\{\partial^2 \ln p(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{b}} \mid \mathbf{y}, \hat{\mathbf{D}}, \mathbf{\Gamma}^{-1} = \mathbf{0}) / \partial(\boldsymbol{\alpha}, \mathbf{b}) \partial(\boldsymbol{\alpha}, \mathbf{b})^T\}^{-1} \tag{3.8}$$

where $\hat{\mathbf{D}}$ maximizes (3.3). Since $p(\boldsymbol{\alpha}, \mathbf{b} | \mathbf{y}, \hat{\mathbf{D}}, \mathbf{\Gamma}^{-1} = 0)$ is proportional to the product of $f(\mathbf{y} | \boldsymbol{\alpha}, \mathbf{b})$ and $p(\mathbf{b} | \hat{\mathbf{D}})$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\mathbf{b}}$ can be calculated using standard optimization routines similar to those used for ordinary maximum likelihood. Further $\hat{\Sigma}_{\alpha b}^{-1}$ is available in closed form.

This approach to approximating posterior means and variances was used by Leonard (1975) and Laird (1978) in work on contingency tables with random parameters. Laird and Louis (1982) consider this approximation in a more general setting.

To estimate $\mathbf{D}$, we must deal with the numerical intractability of the marginal likelihood

(3.3). Leonard and Laird used different approaches to the estimation of variance components, either of which could be applied here. Leonard suggests using a prior with known parameters for the components of **D**, then using joint posterior modes to estimate simultaneously $\boldsymbol{\alpha}$, **b** and **D**. Dempster, Selwyn and Weeks (1983) use a similar approach. Laird (1978) takes the maximization of (3.3) as the objective, but approximates its derivatives in calculating $\hat{\mathbf{D}}$. This approximation, which relies on the special form of the derivatives, is discussed in §4.

4. Computation via the EM Algorithm

The EM algorithm is a general-purpose optimization routine for computing maximum likelihood estimates in an incomplete data setting. In their paper characterizing the algorithm, Dempster, Laird and Rubin (1977) give empirical Bayes and variance-components problems as two examples of its application. In the longitudinal-data setting, we regard the individual parameters, **b**, as missing data, thereby invoking the incomplete-data setting.

Suppose that **b** were observed. Then the MLE of **D** would be

$$\hat{\mathbf{D}} = \sum_{1}^{N} \mathbf{b}_i\mathbf{b}_i^T/N. \tag{4.1}$$

If **D** has a special form (diagonal, for example), (4.1) simplifies, but the components of $\hat{\mathbf{D}}$ remain quadratic forms in **b**. Equation (4.1) follows from the fact that, if **b** were observed, its likelihood would have the exponential-family form with sufficient statistic $\sum_1^N \mathbf{b}_i\mathbf{b}_i^T$. Equation (4.1) thus defines the M-step for each iteration of the EM algorithm.

The E-step of each iteration is based on calculating the expected value of the sufficient statistic, conditional on the observed data vector, **y**, and the current value of $\hat{\mathbf{D}}$. Thus calculation of

$$\hat{\mathbf{t}} = \sum_{1}^{N} \mathrm{E}(\mathbf{b}_i\mathbf{b}_i^T \mid \mathbf{y}_i, \hat{\mathbf{D}}, \boldsymbol{\Gamma}^{-1} = \mathbf{0}) \tag{4.2}$$

defines the E-step of each iteration.

Exact calculation of (4.2) requires N numerical integrations, each in k dimensions. However, if we use the normal approximation proposed in §3, calculation of (4.2) is straightforward. Here we assume that the conditional distribution of **b**, given **y**, is approximately normal with mean $\hat{\mathbf{b}}$ and covariance matrix $\hat{\Sigma}_b$, as defined in (3.7) and (3.8).

Computation of $\hat{\mathbf{b}}$ as defined in (3.7) will generally require iteration, so the algorithm which we suggest requires an iterative procedure (Newton–Raphson) at each E-step of the EM algorithm. We have developed software in Fortran to implement this approach to calculating $\hat{\alpha}$, $\hat{\mathbf{b}}$, $\hat{\Sigma}_\alpha$, $\hat{\Sigma}_b$ and $\hat{\mathbf{D}}$. Our experience using the data set described in §6 suggests that the method may be prohibitively expensive for exploratory analyses when all the n_i and k are large, even though N is relatively small. Savings could possibly be realized by using only one Newton step to approximate (4.2) at each E-step of the algorithm. Alternatively, if each n_i is large and k is small, it may be more efficient to use numerical integration to approximate (4.2) at each E-step.

5. A Two-Step Approach to Estimation

To deal with the analytic intractability of the marginal likelihood, Korn and Whittemore (1979) proposed a method for estimating $\boldsymbol{\alpha}$ and **D** which we call two-step (TS) since it involves two separate estimation problems. First, logistic regression parameters are estimated for each individual. Then asymptotic ML theory is used whereby the estimated parameters are assumed to be approximately normal with means equal to the true individual parameters and covariance matrices equal to the inverse of the observed information matrices. These sampling variances are subsequently treated as known. Since the true parameters are assumed normally distributed at Stage 2, standard normal theory can now be used to estimate $\boldsymbol{\alpha}$ and $\mathbf{D}$ (Korn and Whittemore, 1979). By definition of the procedure, it is applicable only when we assume that the growth-curve model holds (i.e. $\mathbf{X}_i = \mathbf{W}_i\mathbf{Z}_i$).

Formally, we define the TS estimators of $\boldsymbol{\alpha}$ and $\mathbf{D}$ as follows. Let $\hat{\boldsymbol{\nu}}_i$ denote the MLE based on (2.1) and let $\mathbf{S}_i$ denote its large-sample variance–covariance matrix. Assume that $\hat{\boldsymbol{\nu}}_i \sim$ MVN $(\boldsymbol{\nu}_i, \mathbf{S}_i)$, and $\boldsymbol{\nu}_i \sim$ MVN$(\mathbf{W}_i\boldsymbol{\alpha}, \mathbf{D})$. Then marginally the $\hat{\boldsymbol{\nu}}_i$ are independently distributed as MVN $(\mathbf{W}_i\boldsymbol{\alpha}, \mathbf{S}_i + \mathbf{D})$, where the $\mathbf{S}_i$ are known. Standard multivariate-normal theory can be used to derive estimates for $\boldsymbol{\alpha}$ and $\mathbf{D}$. An REML version using a flat prior on $\boldsymbol{\alpha}$ is a straightforward extension. Korn and Whittemore also suggested a simplified two-step method (STS) whereby each component of $\boldsymbol{\nu}_i$ is analyzed separately. This is equivalent to setting all the off-diagonal terms to zero in $\mathbf{D}$ and in each $\mathbf{S}_i$.

The two-step approach has much to recommend it, since analytically it is far easier and cheaper to implement. However, if individual response rates are low, or if n_i is small relative to k, then some of the $\hat{\nu}_{ij}$ may be $\pm\infty$. These individuals must be excluded from the analysis. This form of case exclusion is far from random and the resulting inferences must be interpreted carefully. Even if no cases are excluded, each n_i should be sufficiently large to ensure that the asymptotic normality assumption for $\hat{\boldsymbol{\nu}}_i$ is valid.

6. Example

This section compares the results of several analyses by the empirical Bayes (EB) and two-step (TS and STS) methods of a panel study of asthmatics described by Korn and Whittemore (1979). The panel consisted of 64 asthmatics living in Garden Grove, California. Daily observations of air pollution, weather conditions and the presence or absence of an asthma attack for each participant were recorded from 13 November 1974 to 29 June 1975. The first 14 days of data were excluded from the analysis. Some data on asthma attacks are missing. Between-individual covariates include sex, age and history of hay fever.

Parameter estimates were computed for all three methods by the EM algorithm. The procedure for the two-step methods was based on Laird and Ware (1982). Our results using the STS method were not identical to the results reported by Korn and Whittemore, apparently because of differences between the data set used in their original analysis and the data set provided to us. Because it proved difficult to resolve the issue of data comparability, we chose to apply the Korn and Whittemore method to our data set and restrict comparisons to results obtained in our reanalysis. In comparing the results for different methods, we focus on the coefficient of Total Suspended Particulates (TSP), a measure of the weight of suspended particles per unit volume of outdoor air. It is the parameter of greatest substantive interest and its estimate is sensitive to the method of analysis.

Korn and Whittemore considered two sets of within-individual covariates in their analysis. The larger model included 11 time-varying covariates: A Markov covariate indicating an asthma attack on the previous day, daily average concentration of TSP (in μg/m^3), minimum daily temperature (in degrees Fahrenheit), humidity (%), 'study period', and six indicator covariates for the days of the week (Monday through Saturday). The 'study period' covariate was an indicator for the first 46 days of the study. Korn and Whittemore's analysis required the elimination of approximately one-third of the available subjects due to the absence of attacks on particular days of the week, or during the first 46 days of the study, and the resulting inability to estimate logistic regression coefficients. To increase the effective sample size, they performed a second analysis utilizing only five

Table 1
Estimated mean, standard error and random-effect standard deviation for five regression coefficients in the 11*-covariate model* ($N = 40$)

Method	Variable				
	Markov	TSP*	Temp.*	Humidity*	Saturday
Estimated mean					
STS	2.46	1.55	−3.37	−3.73	0.55
TS	2.18	0.67	−1.72	−4.56	0.55
EB	2.41	1.08	−2.75	−4.70	0.54
Standard error					
STS	0.23	0.95	8.60	3.42	0.17
TS	0.18	0.91	7.63	3.16	0.16
EB	0.21	0.88	7.69	3.22	0.17
Standard deviation of random effect					
STS	1.31	2.61	35.73	9.83	0.55
TS	0.99	2.34	28.43	7.52	0.56
EB	1.23	2.73	32.24	10.00	0.68

* $\times 10^3$.

within-individual covariates. This was accomplished by eliminating the 'study period' covariate and collapsing the six day-of-week covariates into a single indicator variable identifying the first half of the week (Sunday through Wednesday).

In our data set, individual logistic-regression parameters of the 11-covariate model were successfully estimated for 40 subjects. We performed the TS, STS and EB analysis for this set of 40 individuals. Estimates of regression coefficients, standard errors and associated variance components for five variables by the three methods are displayed in Table 1. While the three methods gave somewhat different estimates of the TSP coefficient, the agreement was quite good for other regression parameters (relative to their standard errors). Neither the TS nor the EB method suggests much of an effect of TSP. The STS method suggests a marginally significant effect.

When the STS and TS methods were implemented for the five-covariate model, logistic regression coefficients were obtained for 54 individuals. Hence, the five-covariate model was fitted for $N = 54$ by all three methods. The five-covariate model was also fitted for $N = 40$ by all three methods, and by the EB method for $N = 64$, to assess the impact of

Table 2
Estimated mean (with standard error) and standard deviation, d, of the random-effect distribution for the TSP regression coefficient, for two models and three sample sizes ($\times 10^3$)

Method	$N = 40$		$N = 54$		$N = 64$	
	Mean (SE)	*d*	Mean (SE)	*d*	Mean (SE)	*d*
11-covariate model						
STS	1.55 (.95)	2.61	—	—	—	—
TS	0.67 (.91)	2.34	—	—	—	—
EB	1.08 (.88)	2.73	0.70 (.81)	2.69	0.86 (.78)	2.68
Five-covariate model						
STS	2.35 (.86)	2.53	1.66 (.81)	2.70	—	—
TS	1.78 (.84)	2.33	1.21 (.78)	2.51	—	—
EB	1.83 (.83)	2.42	1.21 (.77)	2.67	1.35 (.75)	2.65

Table 3
Crude asthma attack rates (per 1000) by period and pollution level

Pollution level	Period	
	1	2
Low	266	224
Medium	272	237
High	278	249

model and sample-size reduction on parameter estimates. Resulting estimates for TSP parameters are shown in Table 2.

The mean regression coefficient for TSP is larger for the five- than for the 11-covariate model for all sample sizes. We attribute this to residual confounding induced by dropping the period and day-of-week variables. The coefficients are consistently smaller for $N = 54$ than for $N = 40$. This underscores the probable bias induced by dropping subjects for whom the logistic model cannot be fitted. The results obtained by the EB method were quite similar for $N = 54$ and $N = 64$, since the additional 10 subjects provided relatively little data. The estimated standard errors and random-effect standard deviations were insensitive to both model specification and sample size, except for the expected decrease in standard error with increasing sample size.

To further investigate the possibility of residual confounding when day of the week and period were eliminated from the model, we computed the mean TSP level and crude asthma attack frequency by period and by day of the week. The comparison of the two periods showed that mean TSP levels were 118.5 $\mu g/m^3$ for Period 1 and 73.7 $\mu g/m^3$ for Period 2, a substantial difference. When days were classified as low (0–55 $\mu g/m^3$), moderate (56–90 $\mu g/m^3$), or high (>90 $\mu g/m^3$) pollution, crude attack frequencies by period and pollution level showed a substantial period effect and trends consistent with a pollution effect within period (Table 3). These results indicate that the elimination of period from the regression model would produce positive confounding bias in the estimated coefficient of TSP.

Similarly, the mean TSP levels by day of the week (beginning with Sunday) were 223, 239, 231, 239, 266, 259 and 257 $\mu g/m^3$, that is, a tendency toward higher readings on Thursday through Saturday. When asthma attack rates were computed by day of the week and pollution level, a corresponding pattern of higher asthma attack frequency on Thursday, Friday and Saturday was observed (Table 4). Thus, day of the week is also a potential confounding variable, but it would appear that most of this confounding would be controlled by the single indicator variable used in the reduced model.

These results demonstrate that, in this setting, the two methods of approximation lead to very similar parameter estimates for a common model and a common data set. They also show that TS methods can be substantially affected by the model and sample-size

Table 4
Crude asthma attack rates (per 1000) by day of week and pollution level

Pollution level	Day of week						
	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
Low	205	248	233	228	241	240	228
Medium	231	229	228	219	257	247	260
High	238	231	233	246	291	287	270

Table 5
Estimated mean (with standard error) and standard deviation, d, of the random-effect distribution for the TSP regression coefficient in the six-covariate model including history of hay fever

Method	$N = 54$		$N = 64$	
	Mean (SE)	d	Mean (SE)	d
Main effect of TSP				
STS	1.06 (1.13)	2.69	—	—
TS	0.56 (1.09)	2.28	—	—
EB	0.74 (1.10)	2.66	0.78 (1.08)	2.64
Interaction of TSP and hay fever				
STS	1.22 (1.62)		—	
TS	1.41 (1.55)		—	
EB	1.04 (1.54)		1.15 (1.50)	

reductions required to achieve estimability. The advantage of the EB approach is that all data and covariates can be used, if desired.

Since Korn and Whittemore found that study participants with a history of hay fever had lower TSP coefficients than those without such a history, a second model including hay fever and its interactions with each time-varying covariate was investigated. To reduce computing costs, which can be substantial for these analyses, the model with five time-varying covariates was chosen for this comparison, even though it appears to introduce some confounding.

The results for all three methods based on 54 subjects and for the EB analysis of 64 subjects showed relatively good agreement for the estimated effect of TSP and its interaction with hay fever (Table 5). Since the interaction term was coded as 1 for participants with no history of hay fever and 0 for those with a positive history, these results indicate a larger regression coefficient for TSP among participants without hay fever. However, the interaction term did not achieve statistical significance in any instance. The parameter estimates also vary appreciably among the three methods.

7. Discussion

These results show that the logistic-linear model can be implemented in a large data set using the approximate empirical Bayes estimation methods described in §3 and §4. When the analysis was restricted to the subjects for whom a given logistic regression model can be fitted by two-step methods, the EB methods gave results similar to those obtained by the TS and STS methods. However, the EB method requires neither model simplification nor elimination of subjects. Since model simplification can introduce confounding, and elimination of subjects with low event rates can also introduce bias, the ability to utilize all available data is of practical significance. These features are especially important when each individual's observation vector is short, since two-step methods fail in that setting.

Estimates reported in Tables 1–3 were obtained with an algorithm that terminated when an iteration produced changes in parameter estimates uniformly less than 10^{-3}. When results for the STS method were compared with values reported by Korn and Whittemore, we noted a substantial difference in the estimated standard deviations of the distributions of the TSP coefficients. Whereas the estimated standard deviation was consistently close to 2.5 in our analyses, Korn and Whittemore reported an estimated standard deviation of zero. To investigate this issue, we repeated the STS analysis for the five-covariate model,

$N = 40$, using both the EM and Newton–Raphson algorithms and a convergence criterion of 10^{-6}. The EM algorithm required several hundred iterations and produced an estimated variance component about one order of magnitude smaller than the value reported in Table 2, along with estimates of the mean regression coefficient and standard error within 1% of the values in Table 2. The Newton–Raphson algorithm produced similar results. This experience conforms with the well-known difficulties encountered by most estimation schemes when the likelihood is maximized for variance estimates at or near zero. Since the estimates of the regression coefficients and their standard errors were insensitive to the convergence criterion, more extensive study of the likelihood surface near $D = 0$ was not undertaken for the present paper.

Acknowledgements

The authors wish to thank the referees for a number of suggestions that led to significant improvement of this paper. Support for this research was provided by Grant GM29745 from the National Institutes of Health.

Résumé

Cet article présente un modèle mixte général pour l'analyse de séries de réponses dichotomiques proposées par un panel. On suppose que les séries de réponses de chaque sujet sont issues d'un modèle logistique dont les coefficients de régression dépendent du sujet et que les paramètres de régression sont distribués normalement dans la population. L'inférence est basée sur l'estimation du maximum de vraisemblance des effets fixés et des composantes de la variance, et l'estimation par la méthode du Bayesien empirique des effets aléatoires. Les solutions exactes sont inaccessibles analytiquement ou par le calcul, mais on propose une approximation utilisant le mode de la distribution a posteriori des effets aléatoires, mise en œuvre au moyen de l'algorithme EM. Cette méthode approchée est comparée avec la méthode plus simple à deux pas proposée par Korn et Whittemore (1979, *Biometrics* **35**, 795–804) sur des données d'un panel d'asthmatiques décrit dans cet article. Un avantage de la stratégie d'estimation proposée ici est la possibilité d'utiliser toutes les données, y compris celles des sujets pour lesquels il n'existe pas suffisamment de résultats pour ajuster un modèle de régression logistique, comme l'exige la méthode de Korn et Whittemore. Cependant, la nouvelle méthode nécessite beaucoup de calculs.

References

Bock, R. D. (1975). *Multivariate Statistical Methods in Behavioral Research.* New York: McGraw-Hill.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete observations. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Dempster, A. P., Selwyn, M. R. and Weeks, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *Journal of the American Statistical Association* **78**, 221–227.

Goldstein, H. (1975). *The Design and Analysis of Longitudinal Studies.* New York: Academic Press.

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

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–340.

Hasselblad, V. (1978). Comparison of methods for the analysis of panel studies. Report No. EPA-600/1-78-043. Research Triangle Park, North Carolina: United States Environmental Protection Agency.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H. and Lehman, R. G. (1977). A general methodology for the analysis of repeated measurement of categorical data. *Biometrics* **33**, 133–158.

Korn, E. L. and Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* **35**, 795–804.

Laird, N. M. (1978). Bayesian estimation methods for two-way contingency tables. *Biometrika* **65**, 581–590.

Laird, N. M. and Louis, T. A. (1982). Approximate posterior distributions for incomplete data problems. *Journal of the Royal Statistical Society, Series B* **44,** 190–200.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38,** 963–974.

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B* **37,** 23–37.

Markus, G. B. (1979). *Analyzing Panel Data.* Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-018. Beverly Hill and London: Sage Publications.

Nesselroade, J. R. and Baltes, P. B. (eds) (1979). *Longitudinal Research in the Study of Behavior and Development.* New York: Academic Press.

Rao, C. R. (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics* **31,** 545–554.

Stebbings, J. H. and Hayes, C. G. (1976). Panel studies of acute health effects of air pollution. *Environmental Research* **11,** 89–111.

Strenio, J. F., Weisberg, H. I. and Bryk, A. S. (1983). Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics* **39,** 71–86.

Ware, J. H. and Wu, M. (1981). Tracking: the prediction of future and missing values from a series of measurements. *Biometrics* **37,** 427–438.

Whittemore, A. S. and Korn, E. L. (1980). Asthma and air pollution in the Los Angeles area. *American Journal of Public Health* **70,** 687–696.

Received January 1982; *revised September* 1983 *and January* 1984