Journal of Consulting and Clinical Psychology
1994, Vol. 62, No. 2, 285-296

0022-006X/94/$3.00

Application of Random-Effects Probit Regression Models

Robert D. Gibbons and Donald Hedeker

A random-effects probit model is developed for the case in which the outcome of interest is a series of correlated binary responses. These responses can be obtained as the product of a longitudinal response process where an individual is repeatedly classified on a binary outcome variable (e.g., sick or well on occasion *t*), or in "multilevel" or "clustered" problems in which individuals within groups (e.g., firms, classes, families, or clinics) are considered to share characteristics that produce similar responses. Both examples produce potentially correlated binary responses and modeling these person- or cluster-specific effects is required. The general model permits analysis at both the level of the individual and cluster and at the level at which experimental manipulations are applied (e.g., treatment group). The model provides maximum likelihood estimates for time-varying and time-invariant covariates in the longitudinal case and covariates which vary at the level of the individual and at the cluster level for multilevel problems. A similar number of individuals within clusters or number of measurement occasions within individuals is not required. Empirical Bayesian estimates of person-specific trends or cluster-specific effects are provided. Models are illustrated with data from mental health research.

There has been considerable interest in random-effects models for longitudinal and hierarchical, clustered, or multilevel data in the statistical literatures for biology (Jennrich & Schluchter, 1986; Laird & Ware, 1982; Ware, 1985; Waternaux, Laird, & Ware, 1989; Hedeker & Gibbons, 1994), education (Bock, 1989; Goldstein, 1987), psychology (Bryk & Raudenbush, 1987; Willett, Ayoub, & Robinson, 1991), biomedicine (Gibbons, Hedeker, Waternaux, & Davis, 1988; Hedeker, Gibbons, Waternaux, & Davis, 1989; Gibbons et al., 1993), and actuarial and risk assessment (Gibbons, Hedeker, Charles, & Frisch, in press). Much of the work cited here has been focused on continuous and normally distributed response measures. In contrast, there has been less focus on random-effects models for discrete data. Gibbons & Bock (1987) have developed a random-effects probit model for assessing trend in correlated proportions, and Stiratelli, Laird, and Ware (1984) have developed a random-effects logit model for a similar application. Using quasi-likelihood methods in which no distributional form is assumed for the outcome measure, Liang and Zeger (1986; Zeger & Liang, 1986) have shown that consistent estimates of regression parameters and their variance estimates can be obtained regardless of the time dependence. Koch, Landis, Freeman, Freeman, and Lehnen (1977) and Goldstein (1991) have illustrated how random effects can be incorporated into log-linear models. Finally, generalizations of the logistic regression model in which the values of all regression coefficients vary randomly over individuals have also been proposed by Wong and Mason (1985) and Conoway (1989).

Robert D. Gibbons and Donald Hedeker, Biometric Lab, University of Illinois at Chicago.

This study was supported by Grant R01 MH 44826-01 A2 from the National Institute of Mental Health Services Research Branch.

Correspondence concerning this article should be addressed to Robert D. Gibbons, Biometric Lab (M/C 913), University of Illinois, 912 South Wood Street, Chicago, Illinois 60612.

The purpose of this article is to describe the random-effects probit model of Gibbons and Bock (1987) and further generalize it for application to a wider class of problems commonly encountered in the behavioral sciences, including hierarchical or clustered samples, estimation of time-varying and time-invariant covariates, marginal maximum likelihood estimation of structural parameters, and empirical Bayesian estimation of person-specific or cluster-specific effects and illustrate its application. A detailed description of data giving rise to the need for this type of statistical modeling is now presented.

Longitudinal Data

In general, we consider the case in which the same units are repeatedly sampled at each level of an independent variable and classified on a binary outcome. Specifically, we are interested in repeated classification of individuals on a series of measurement occasions over time. In a clinical trial, for example, patients may be randomly assigned to treatment and control conditions and repeatedly classified in terms of presence or absence of clinical improvement, side effects, or specific symptoms. We also may be interested in comparing rate of improvement (e.g., proportion of patients displaying the symptom) between treatment and control conditions (Gibbons & Bock, 1987), which may be tested by assuming that each individual follows a straight-line regression on time. Probability of a positive response depends on that individual's slope and a series of covariates that may be related to the probability of response. Covariates can take on fixed values for the length of the study (e.g., sex or type of treatment) or occasion-specific values (e.g., social supports or plasma level of a drug). Table 1 illustrates the longitudinal data case.

Table 1 describes a 4-week longitudinal clinical trial in which patients who are randomly assigned to one of two treatment groups (e.g., active treatment versus placebo control) are repeatedly classified on a binary outcome measure. Drug plasma lev-

Table 1
Longitudinal Data

	Time ($k = 0, \ldots, n_i - 1$) at week:				
Subject ($i = 1, \ldots, N$)	0	1	2	3	4
Outcome_1	0	0	1	1	0
Treatment group_1	1	1	1	1	1
Plasma_1	10	10	20	30	10
Outcome_2	1	0	0	*na*	0
Treatment group_2	0	0	0	*na*	0
Plasma_2	30	20	20	*na*	20
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
Outcome_N	0	0	0	1	1
Treatment group_N	1	1	1	1	1
Plasma_N	20	30	40	40	10

Note. $i = 1 \ldots N$ subjects; $k = 1 \ldots n_i$ observations on subject i; *na* = not available.

els are included to determine if a relationship exists between blood level and clinical response. Note that treatment group is constant over time (time-invariant covariate), whereas plasma level is measured in time-specific values (time-varying covariate). There is no requirement that each individual must have measurements on each occasion. Indeed, Subject 2 appears to have been unavailable for the Week 3 assessment. The model provides flexible treatment of missing data. It assumes that available data accurately represent trend. For example, if a subject drops out because of nonresponse, we assume that absence of positive trend will persist as if the patient had remained in the study. More generally, we assume that available data characterize the deviation of each subject from the group-level response.

Clustered Data

An analogous situation to the longitudinal data problem arises in the context of clustered data. Here, repeated classifications are made on individual members of the cluster. To the extent that classifications between members of a cluster are similar (i.e., intraclass correlation), responses are not independent and the assumptions of typical models for analysis of binary data (e.g., log-linear models, logistic regression, chi-square statistics) do not apply. These methods assume there are n independent pieces of information, but to the extent that intraclass correlation is greater than zero, this is not true.

As in the longitudinal data case, we can have covariates at two levels. Person-specific and cluster-specific covariates can be simultaneously estimated. In addition, there is no requirement that clusters have the same number of members. As an example, consider a family study in which presence or absence of depression in each member is evaluated in terms of overall level of familial support, life events, sex, and age. The family represents the cluster level variable and familial support is a cluster-level covariate. Life events, sex, and age vary at the individual level within a familial cluster. Because families vary in size, there is no restriction that number of members in a familial cluster be constant. We must be careful to include intrafamilial correlation of these classifications in our computations (i.e., siblings are more likely to exhibit comorbidity for depression than unrelated individuals). Treatment of these data as if they were independent (i.e., from unrelated individuals) would result in overly optimistic (i.e., too small) estimates of precision (i.e., standard errors). Had we examined proportion of affected relatives, we would have lost the ability to correlate outcome with individual personal characteristics (i.e., sex and life events of the relative). Considerable statistical power is gained if the unique portion of each individual's response is included in the analysis.

An example of a clustered dataset is presented in Table 2. The data presented in Table 2 may apply to the family study, where clusters represent N families each with n_i members. For each family, there is a cluster-level covariate (i.e., family support) and an individual-level covariate (i.e., age of the relative). These data may be collected to examine effects of age and family support on incidence of depression (i.e., outcome) within families.

A Random-Effects Probit Regression Model

Gibbons and Bock (1987) have presented a random-effects probit regression model to estimate trend in a binary variable measured repeatedly in the same subjects. In this article, we provide an overview of a general method of parameter estimation for both random and fixed effects. We also discuss empirical Bayes estimates of person-specific or cluster-specific effects and corresponding standard errors, so that trend at group and individual levels may be evaluated. In the first two sections, we describe a model with one random effect, adaptable to either clustered or longitudinal study designs. In the third section, we describe a model with two random effects suited to longitudinal data analysis.

A Model for Clustered Data

We begin with the following model for subject k (where $k = 1, 2, \ldots, n_i$) in cluster i ($i = 1, \ldots, N$ clusters in the sample).

Table 2
Clustered Data

	Subject ($k = 1, \ldots, n_i$):		
Cluster ($i = 1, \ldots, N$)	1	2	$3 \ldots n_i$
Outcome_1	0	0	1 ... 0
Family support_1	1	1	1 ... 1
Age_1	10	10	20 ... 10
Outcome_2	1	0	0 ... 0
Family support_2	0	0	0 ... 0
Age_2	30	20	20 ... 20
...	...	...	...
...	...	...	...
...	...	...	...
Outcome_N	0	0	0 ... 1
Family support_N	1	1	1 ... 1
Age_N	20	30	40 ... 10

Note. $i = 1 \ldots N$ clusters; $k = 1 \ldots n_i$ subjects in cluster i.

$$y_{ik} = \alpha_i + \beta_1 x_{1_i} + \beta_2 x_{2_{ik}} + \epsilon_{ik}, \quad (1)$$

where y_{ik} = the unobservable continuous "response strength" for subject k in cluster i; α_i = the random effect of cluster i; β_1 = the fixed effect of the cluster level covariate x_{1_i}; β_2 = the fixed effect of the subject level covariate $x_{2_{ik}}$; and ϵ_{ik} = an independent residual distributed N(0, σ^2). Here, α_i represents a coefficient for the random cluster effect. The assumed distribution for the α_i is N(μ_α, σ_α^2). The coefficient α_i represents the deviation of cluster i from the overall population mean μ_α conditional on the covariate values for that person and cluster. Conditional on the covariates, the $n_i \times 1$ vector of subject response strengths for cluster i, $\mathbf{y}_i$ are multivariate normal with mean $E(\mathbf{y}_i) = \mathbf{1}_i\mu_\alpha + \mathbf{X}_i\beta$; and covariance matrix $V(\mathbf{y}_i) = \sigma_\alpha^2\mathbf{1}_i\mathbf{1}_i' + \sigma^2\mathbf{I}_i$, where $\mathbf{1}_i$ is an $n_i \times 1$ unity vector, $\mathbf{I}_i$ is an $n_i \times n_i$ identity matrix, β is the $p \times 1$ vector of covariate coefficients, and $\mathbf{X}_i$ is the $n_i \times p$ covariate matrix.

To relate the manifest dichotomous response with the underlying continuous response strength y_{ik}, Gibbons and Bock (1987) used a "threshold concept" (Bock, 1975, p. 513). They assume the underlying variable is continuous, and that in the binary response setting, one threshold value (γ) exists on the continuum of this variable. The presence or absence of a positive response for subject k in cluster i is determined by whether underlying response strength exceeds the threshold value. When response strength exceeds the threshold, a positive response is given (coded $v_{ik} = 1$), otherwise a negative response is given (coded $v_{ik} = 0$).

Using the threshold model we can express probability of a positive response in terms of the value $1 - \Phi(z_k)$; that is, the area under the standard normal distribution function at the point z_k, where z_k is the normal deviate given by $(\alpha + \beta_1 x_1 + \beta_2 x_{2_k} - \gamma)/\sigma$. Additionally, origin and unit of z may be chosen arbitrarily, so for convenience, let $\sigma = 1$ and $\gamma = 0$. Probability of a particular pattern of responses for the n_i subjects in cluster i, denoted $\boldsymbol{v}_i$, is the product of probabilities for the n_i binary responses, namely,

$$l(\boldsymbol{v}_i \mid \alpha, \beta) = \prod_{k=1}^{n_i} [\Phi(z_{ik})]^{1-v_{ik}}[1 - \Phi(z_{ik})]^{v_{ik}}. \quad (2)$$

Thus, marginal probability of this pattern is given by.

$$h(\boldsymbol{v}_i) = \int_\alpha l(\boldsymbol{v}_i \mid \alpha, \beta)g(\alpha)d\alpha,$$

where $l(\boldsymbol{v}_i \mid \alpha, \beta)$ is given above, and $g(\alpha)$ represents the distribution of α in the population, (normal distribution with mean μ_α and variance σ_α^2).

Orthogonalization of the Model Parameters

In parameter estimation for the random-effects probit regression model, Gibbons & Bock (1987) orthogonally transform the response model to use the marginal maximum likelihood estimation procedure for the dichotomous factor analysis model discussed by Bock and Aitken (1981). The orthogonalization can be achieved by letting $\alpha = \sigma_\alpha\theta + \mu_\alpha$, where σ_α is the standard deviation of α in population. Then $\theta = (\alpha - \mu_\alpha)/\sigma_\alpha$, and so, $E(\theta) = 0$ and $V(\theta) = \sigma_\alpha^{-1}\sigma_\alpha^2\sigma_\alpha^{-1} = 1$. The reparameterized model is then written as

$$z_k = \mu_\alpha + \sigma_\alpha\theta + \boldsymbol{x}_k'\beta \quad (3)$$

and the marginal density becomes

$$h(\boldsymbol{v}_i) = \int_\theta l(\boldsymbol{v}_i \mid \theta, \beta)g(\theta)d\theta, \quad (4)$$

where $g(\theta)$ represents the distribution of the θ vector in the population; that is, the standard normal density. Further details of the marginal maximum likelihood estimation procedure are provided by Gibbons and Bock (1987).

Estimating Cluster-Specific Effects

It may be desirable to estimate level of response strength or propensity for a positive response α_i. A good choice for this purpose (Bock & Aitkin, 1981; Gibbons & Bock, 1987) is the expected a posteriori (EAP) value (Bayes estimate) of θ_i, given the binary response vector $\boldsymbol{v}_i$ and covariate matrix X_i of cluster i.

$$\hat{\theta}_i = E(\theta_i \mid \boldsymbol{v}_i, X_i) = \frac{1}{h(\boldsymbol{v}_i)} \int_{-\infty}^{\infty} \theta_i l(\boldsymbol{v}_i \mid \theta, \beta)g(\theta)d\theta. \quad (5)$$

Similarly, the standard error of $\hat{\theta}_i$, which may be used to express precision of the EAP estimator, is given by

$$\sigma(\hat{\theta}_i \mid \boldsymbol{v}_i, X_i) = \frac{1}{h(\boldsymbol{v}_i)} \int_{-\infty}^{\infty} (\theta_i - \hat{\theta}_i)^2 l(\boldsymbol{v}_i \mid \theta, \beta)g(\theta)d\theta. \quad (6)$$

These quantities can be evaluated using Gauss-Hermite quadrature as described in Gibbons and Bock (1987) or Bock and Aitkin (1981). Estimates of α_i can be recovered by $\alpha_i = \sigma_\alpha\hat{\theta}_i + \hat{\mu}_\alpha$ using the marginal maximum likelihood estimates of the parameters. Because the prior distribution $g(\theta)$ is normal, these linear transformations of EAP estimates are also EAP estimates.

A Model For Longitudinal Data

The model in the previous section can be adapted for longitudinal data as follows. We begin with the model for response on time point k (where $k = 1, 2, \ldots, n_i$) for subject i ($i = 1, \ldots, N$ subjects in the sample):

$$y_{ik} = \alpha_i + \beta_0 + \beta_1 t_{ik} + \beta_2 x_{2_i} + \beta_3 x_{3_{ik}} + \epsilon_{ik}, \quad (7)$$

where y_{ik} = the unobservable continuous "response strength" or "propensity" on time point k for subject i; t_{ik} = is the time (i.e., day, week, year, etc.) that corresponds to the kth measurement for subject i; β_0 = the overall population intercept or response propensity at baseline $t = 0$; β_1 = the overall population trend coefficient describing rate of change in response propensity over time; α_i = the random effect for subject i; β_2 = the fixed effect of the subject level covariate x_{2_i}; β_3 = the fixed effect of the time-specific covariate $x_{3_{ik}}$; and ϵ_{ik} = an independent residual distributed N(0, σ^2). Here, α_i is a coefficient describing deviation of subject i from the overall group response conditional on the covariate vector for that subject. The assumed distribution for α_i is N(μ_α, σ_α^2). In practice, population level intercept β_0 and trend β_1 are incorporated into the model as the first two columns of X_i, where the first column is a vector of ones and the second column contains the n_i measurement occasions for

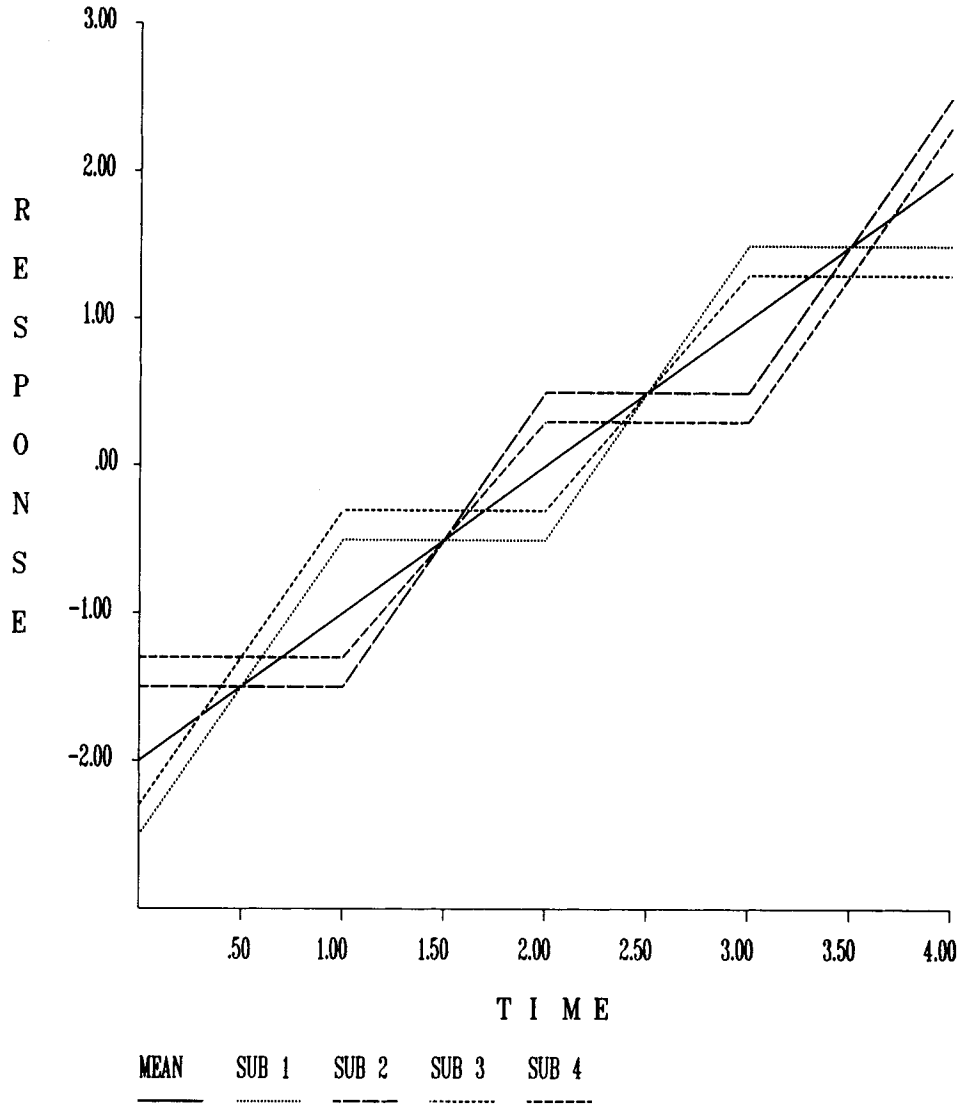

Figure 1. Fixed effects model: time versus response. Average response strength and four typical subjects (SUBs).

subject i (i.e., t_{ik}). Therefore, the same set of likelihood equations and solution derived for the clustered case directly applies also to the longitudinal problem.

Had we ignored the person-specific component of variation in the longitudinal response process and modeled these data with probit or logistic regression analysis using time as the independent variable (i.e., assuming repeated classification were independent), then we would have had to assume deviations in response propensity from the overall group trend vary randomly as well. This assumption, depicted in Figure 1, illustrates that for the fixed-effects model, an individual's deviation from the overall group response propensity may be positive on one occasion and negative on another, an implausible view of the longitudinal response process. Particularly for short-term studies, subjects deviate systematically from the overall group level trend based on measured or unmeasured characteristics that increase or decrease response probability. These characteristics exhibit random variability in the subject population and, to a lesser degree, within an individual over a fixed time.

The model in Equation 7 is termed "random intercept" because person-specific deviations must be parallel to the average trend (see Figure 2). The model is analogous to a mixed-model analysis of variance (ANOVA) for continuous response data. Figure 2 shows that overall response propensity level varies from individual to individual but that deviations from the overall group trend are constant within an individual over time. The model is not plausible for two reasons. First, in many controlled clinical trials, subjects are selected to be similar at baseline but are quite heterogeneous in terms of their response to treatment over time. In this example, it is the trend that is random and not the intercept. Second, in naturalistic studies, for example, many studies of mental health services, there is variability in both the

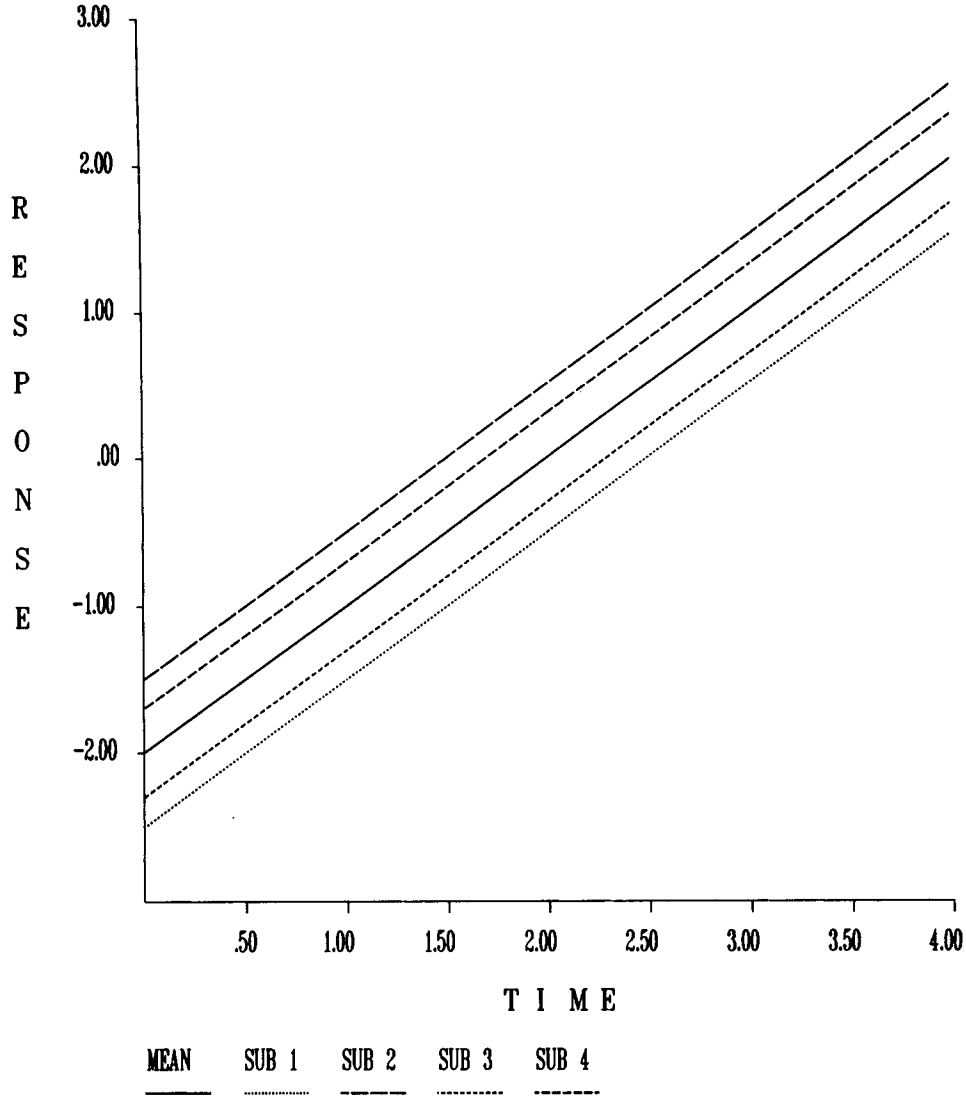

Figure 2. Random intercept model. Average response strength and four typical subjects (SUBs).

intercept (i.e., the subjects are not screened on the basis of severity of illness) and trend (i.e., the efficacy of the services varies greatly in the population of potential recipients). In either case, a model that only allows for person-specific deviations at baseline (i.e., a random intercept model) seems poorly suited for problems in mental health research.

Alternatively, a "random trend" model could be considered as follows.

$$y_{ik} = \beta_0 + \alpha_i t_{ik} + \beta_1 x_{1_i} + \beta_2 x_{2_{ik}} + \epsilon_{ik}, \quad (8)$$

Here, we assume a common intercept or starting point for all subjects (plus or minus random error ϵ_{ik}), and person-specific deviations in the slope of each subject's trend line from the average group trend line. This model is depicted in Figure 3, which shows that deviations from average response rate increase over time, as each subject has an individual rate parameter. Although the solution is similar to the clustered case and random intercept model, the underlying response process assumed by the random trend model is quite different. Modifications to likelihood equations and their solution follow from the derivation given by Gibbons and Bock (1987). In the following section, we consider a model with two random effects, a random intercept and a random trend.

A Model With Two Random Effects

In the previous section, we developed a model with a single random effect. However, both the intercept (i.e., baseline level) and slope (i.e., the rate at which change occurs) can exhibit systematic person-specific deviations from overall population level values (see Figure 4). In this case, the model must be further generalized to the case of two random effects; that is,

$$y_{ik} = \alpha_{0i} + \alpha_{1i} t_{ik} + \beta_1 x_{1_i} + \beta_2 x_{2_{ik}} + \epsilon_{ik}. \quad (9)$$

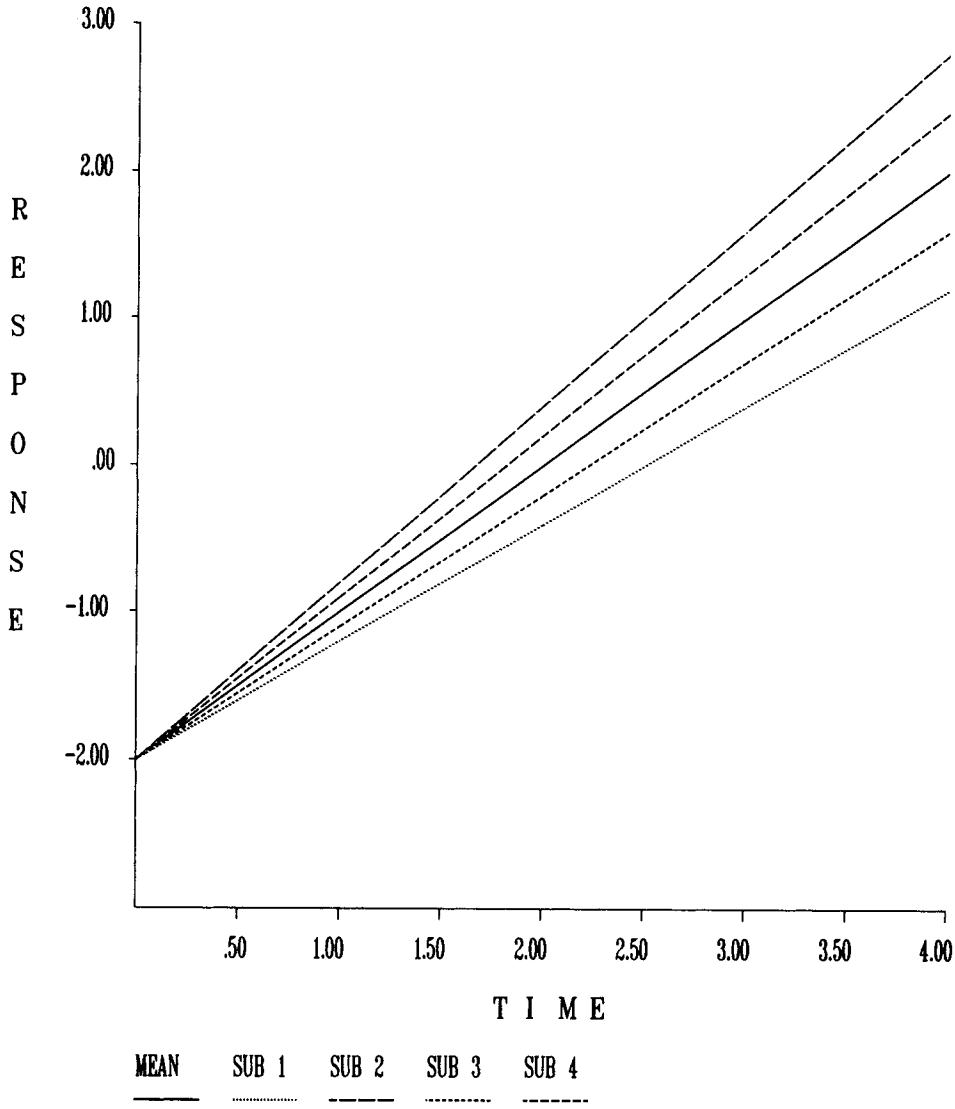

Figure 3. Random trend model. Average response strength and four typical subjects (SUBs).

Here α_{0i} represents the deviation for subject i from the overall group intercept and α_{1i} represents the deviation for subject i from the overall group trend. We assume that distribution of α_0 and α_1 in the population is bivariate normal $N(\mu, \Sigma)$, with

$$\mu = \begin{bmatrix} \mu_{\alpha_0} \\ \mu_{\alpha_1} \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma^2_{\alpha_0} & \sigma_{\alpha_0\alpha_1} \\ \sigma_{\alpha_0\alpha_1} & \sigma^2_{\alpha_1} \end{bmatrix}$$

The model implies that conditional on the covariates, the observations are multivariate normal with mean $E(y) = T\mu$ and covariance matrix $V(y) = T\Sigma T' + \sigma^2 I$, where, for example,

$$T = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & n-1 \end{bmatrix}$$

and σ^2 is a residual variance assumed constant over time. The conditional probability for response pattern of subject i (i.e., v_i) is then

$$l(v_i | \alpha_0, \alpha_1, \beta) = \prod_{k=1}^{n_i} [\Phi(z_{ik})]^{1-v_{ik}}[1 - \Phi(z_{ik})]^{v_{ik}}. \quad (10)$$

where

$$z_{ik} = (\alpha_{0i} + \alpha_{1i} t_{ik} + \beta_1 x_{1i} + \beta_2 x_{2_{ik}} - \gamma)/\sigma.$$

Thus, marginal probability of this pattern is given by

$$h(v_i) = \int_{\alpha_0} \int_{\alpha_1} l(v_i | \alpha_0, \alpha_1, \beta) g(\mu, \Sigma) d\alpha_1 d\alpha_0, \quad (11)$$

where $g(\cdot)$ is the bivariate normal probability density of α_0 and α_1. The method of estimation for the parameters of this model was originally described by Gibbons and Bock (1987); this article contains full details concerning the estimation procedure.

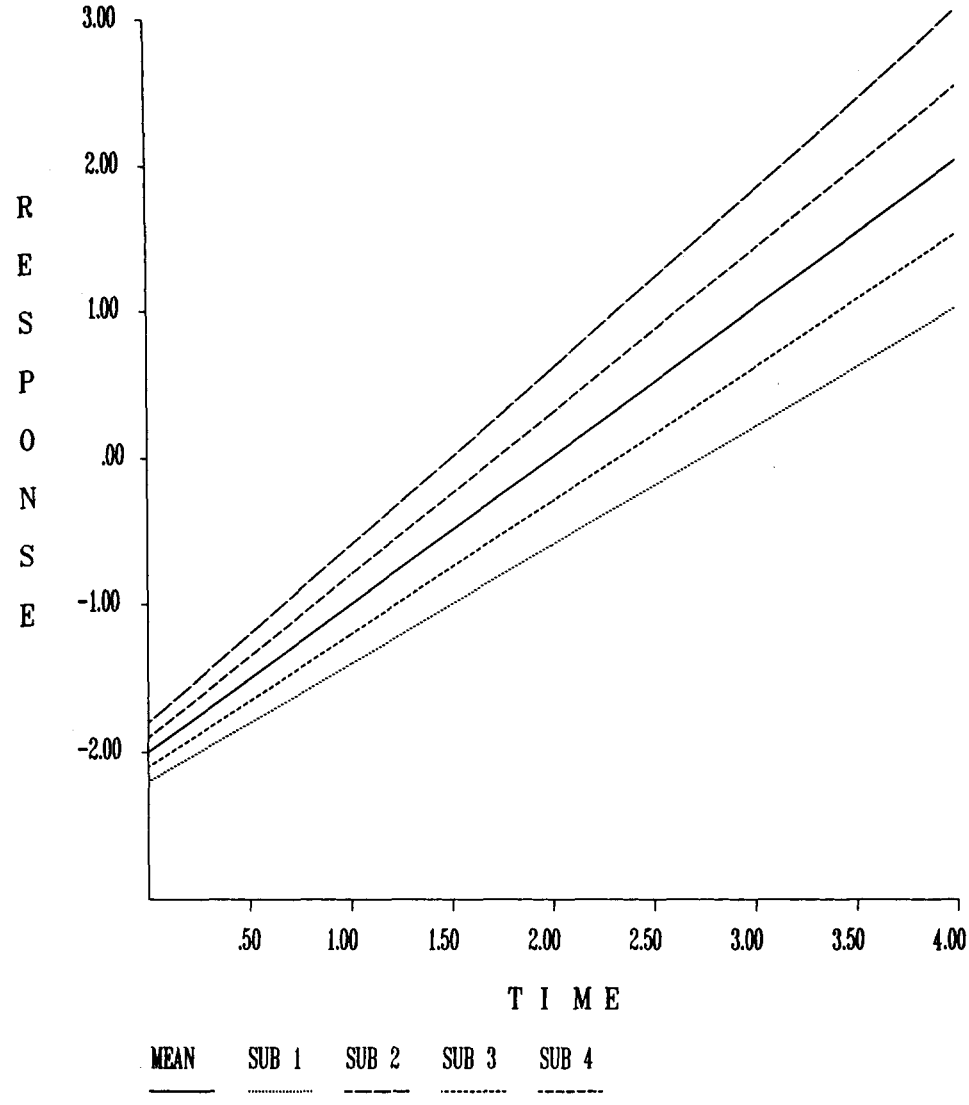

Figure 4. Random intercept and trend model. Average response strength and four typical subjects (SUBs).

Illustration

To illustrate application of the random-effects probit model to clustered and longitudinal data, we examined data from the National Institute of Mental Health Schizophrenia Collaborative Study. Specifically, we examined Item 79 of the Inpatient Multidimensional Psychiatric Scale (IMPS; Lorr & Klett, 1966). Item 79, "Severity of Illness," was originally scored on a 7-point scale ranging from *normal, not at all ill* (1) to *among the most extremely ill* (7). For the purpose of this analysis, we dichotomized the measure between *mildly ill* (3) to *moderately ill* (4). Gibbons, Hedeker, Waternaux, & Davis (1988) analyzed these data in their original metric using a random-effects regression model for continuous response data. Experimental design and corresponding sample sizes are displayed in Table 3.

Table 3 reveals that the longitudinal portion of the study is highly unbalanced. There are large differences in the number of measurements made in the 6 weeks of treatment. In the first analysis, we treat these data as if subjects represent 440 clusters which include one to seven repeated observations, ignoring the longitudinal nature of the repeated observations. Both fixed-effects and random-effects probit models were fitted to these data, using sex and treatment group as covariates. The fixed-effect model allows one to simply ignore the fact that there were repeated measurements from each subject and incorrectly assume that all observations (i.e., both within and across subjects) were independent. Each treatment group was contrasted to the placebo control group. Results are presented in Table 4. The scale of the parameter estimates corresponds to the probit response function (see Finney, 1971). For example, Table 4 reveals that for the random-effects model, the overall response level in placebo patients (i.e., the intercept, because the placebo group was dummy coded as 0 0 0 on the three treatment-related effects) is -1.327 and the total variance is $\sqrt{1 + .555^2}$ (i.e., the

Table 3
Experimental Design and Weekly Sample Sizes

Treatment group	Sample size at week: 0	1	2	3	4	5	6
Placebo	110	108	5	89	2	2	72
Chlorpromazine	110	108	3	96	4	5	87
Fluphenazine	114	108	2	100	2	2	89
Thioridazine	106	107	4	93	3	0	90

residual variance which is fixed at 1 plus the random effect variance of 0.555^2). The corresponding normal probability of illness in the placebo group is given by $\Phi(-1.327/\sqrt{1 + .555^2}) = \Phi(-1.160) = .877$, or an overall estimated proportion of well placebo patients of .123 (i.e., averaging over time). In contrast, the estimated difference between fluphenazine and placebo was 0.834; hence the corresponding probability is given by $\Phi[(-1.327 + .834)/\sqrt{1 + .555^2}] = \Phi(-0.431) = .666$, or an overall estimated proportion of well placebo patients of .334 (i.e., averaging over time). Overall, the model indicates that fluphenazine produces a 21% increase in response relative to placebo (i.e., 33.4% versus 12.3%), when response is defined as mildly ill or better. Of course, we would expect this difference to be larger at the end of treatment as will be shown in the results of the longitudinal analysis.

Table 4 also reveals that addition of the random cluster (i.e., in this case, person) effect is highly significant in the ratio of MLE to SE and in the improvement in fit likelihood ratio chi-square statistic ($\chi^2_1 = 35.02, p < .0001$). As expected, the fixed-effects model considerably underestimated standard errors of parameter estimates. Additionally, and unanticipated, the fixed-effects model also underestimated maximum likelihood parameter estimates. Simply ignoring the within-subject nature of these data is clearly not a good idea. The effect of sex approached significance in the fixed-effects model but was not significant in the random-effects model. All three active treatments

Table 4
Parameter Estimates, Standard Errors, and Probabilities for NIMH Schizophrenia Collaborative Study Clustered Example

Fixed and random effects	Fixed MLE	*SE*	*p*<	Random MLE	*SE*	*p*<
Fixed effects						
Intercept	−1.151	.105	.0001	−1.327	.168	.0001
Sex	.095	.056	.0900	.103	.088	.2440
Chlor vs. Pla	.445	.085	.0001	.521	.135	.0001
Fluph vs. Pla	.726	.083	.0001	.834	.136	.0001
Thior vs. Pla	.521	.084	.0001	.615	.132	.0001
Random effects[a]						
σ_{α_0}				.555	.078	.0001

Note. NIMH = National Institute of Mental Health; MLE = maximum likelihood estimate. Chlor = chlorpromazine; Pla = placebo; Fluph = fluphenazine; Thior = thioridazine.
[a] Log L = −944.65 for fixed effects and −927.14 for the random effects model. For change, $\chi^2 = 35.02, p < .0001$.

exhibited significant improvement relative to placebo controls for both models.

The second analysis accounted for the longitudinal nature of the study design. Here we fitted a fixed-effects model and one- and two-random-effects models to these data. Results of this analysis are presented in Table 5. Table 5 reveals different results depending on whether or not random effects are included. As in the previous analysis, the fixed-effects model underestimates standard errors since it assumes all measurements are independent. Similarly, standard errors for the model with two random effects were equal to or greater than those for the model with one random effect (i.e., random intercept model). The model with one random effect significantly improved fit over the fixed-effects model ($\chi^2_1 = 131.04, p < .0001$), and the model with two random effects significantly improved fit over the model with one random effect ($\chi^2_1 = 73.70, p < .0001$). Person-specific variability in intercepts $\sigma_{\alpha_0} = .860$ and slopes $\sigma_{\alpha_1} = .630$ were both significant ($p < .0001$), but uncorrelated ($\sigma_{\alpha_0\alpha_1} = .056, p < .48$).

Sex, main effect of treatment, and a Treatment × Time interaction were examined. Models with both one and two random-effects revealed significant Treatment × Time interactions for all three active treatments versus placebo control, although magnitude was somewhat greater for the model with two random effects indicating that differences between treatment groups and the placebo control group were linearly increased over the 6-week study. However, the fixed-effects model did not identify significant treatment by time interactions. Only the main effect (i.e., averaging over time points) corresponding to difference between fluphenazine and placebo was significant. In contrast, the fixed-effects model identified a significant sex effect not found in either random-effects model.

These results illustrate that ignoring systematic person-specific effects leads to poor model fit, and can bias the maximum likelihood estimates, standard errors, and probability values associated with tests of treatment-related effects. Indeed, had we naively applied a traditional probit or logistic regression model to these data, we would have incorrectly concluded thioridazine and chlorpromazine did not have any beneficial effects relative to placebo control.

For a better understanding of these differences, predicted probability of illness curves for men and women are displayed in Figures 5 and 6, respectively. The predicted response probabilities are a direct function of the estimated parameter values in Table 5. Comparison of Figures 5 and 6 support the finding of a nonsignificant difference between male and female subjects in that the response curves are quite similar. Similarly, the treatment versus control differences are clearly evident in these figures, with consistent differences emerging as early as 1 week.

Discussion

It should be clear from the material presented that much of the same rich structure that can be extracted from continuous data using random-effects regression models is also available for studies involving binary outcomes. The random-effects probit model with numerical integration presented here is one such model. In contrast to other approaches (e.g., Stiratelli et al., 1984; Wong & Mason, 1985), we restrict the random effects to

Table 5
Parameter Estimates, Standard Errors, and Probabilities for NIMH Schizophrenia Collaborative Study Longitudinal Example

Fixed and random effects	Fixed			1 Random effect			2 Random effects		
	MLE	*SE*	*p*	MLE	*SE*	*p*	MLE	*SE*	*p*
Fixed effects									
Intercept	−1.777	.158	<.0001	−2.630	.314	<.0001	−2.507	.457	<.0001
Slope	.217	.037	<.0001	.309	.042	<.0001	.102	.105	<.33
Sex	.126	.056	<.02	.178	.140	<.20	.215	.188	<.25
Chlor vs. Pla	.265	.175	<.13	.395	.285	<.16	.050	.285	<.86
Fluph vs. Pla	.516	.187	<.006	.810	.303	<.008	.209	.339	<.54
Thior vs. Pla	.314	.170	<.07	.357	.284	<.21	.079	.284	<.78
C vs. Pla × T	.064	.050	<.20	.111	.055	<.04	.427	.136	<.002
F vs. Pla × T	.102	.054	<.06	.164	.061	<.007	.706	.155	<.0001
T vs. Pla × T	.078	.050	<.12	.165	.061	<.007	.526	.144	<.0002
Random effects[a]									
σ_{α_0}				1.180	.112	<.0001	.860	.220	<.0001
$\sigma_{\alpha_0\alpha_1}$							.056	.093	<.48
σ_{α_1}							.630	.112	<.0001

Note. NIMH = National Institute of Mental Health; MLE = Maximum likelihood estimate; Chlor = chlorpromazine; Pla = placebo; Fluph = fluphenazine; Thior = thioridazine.
[a] Log L = −780.81 for the fixed effects model; −715.29 for the model with one random effect, and −678.44 for the model with two random effects. Chi-square values for change are as follows: for the model with one random effect, $\chi^2 = 131.04$, $p < .0001$; for the model with two random effects, $\chi^2 = 73.70$, $p < .0001$.

the intercept and slope of the trend line, treating covariates as fixed. These other approaches typically would treat all estimated coefficients as random. There are advantages and disadvantages to both approaches. With only one or two random effects, the likelihood may be evaluated numerically as presented here. Furthermore, Bock and Aitkin (1981) have shown how the assumption of multivariate normality of the underlying random effect distribution can be relaxed and other distributions can be fitted or nonparametric estimates of the underlying density can be obtained. This generalization is possible here as well but would not be available where the integrals in Equation 4 are approximated by a multivariate normal distribution with the same mode and curvature of the mode as the true posterior (i.e., Bayes modal estimates) as in Stiratelli et al. (1984) and Wong and Mason (1985). Alternatively, as the number of random effects increase beyond three or four, the numerical integration becomes computationally intractable. Anderson and Aitkin (1985) have developed a similar model for examining interviewer variability that also uses numerical integration to obtain maximum likelihood parameter estimates.

Some discussion of missing data is appropriate here. Laird (1988) has described three categories of missing data: missing completely at random; ignorable nonresponse; and nonignorable nonresponse. Although data missing completely at random is easiest to cope with, it is probably not a plausible assumption for longitudinal studies in which subjects often drop out during the course of the study, never to return.

The second category of ignorable nonresponse states that missing data are ignorable as long as they are explained by terms in the model or the available outcome data for each subject. For example, if in a clinical trial patients on placebo drop out more frequently than patients on active treatment, the missing data are ignorable as long as treatment is included as a covariate in the model. As another example, if patients who do poorly during their participation in the study drop out, we expect that they would have continued not to benefit from treatment, hence the distribution of the unobserved outcomes is known conditional on the distribution of the available outcomes (i.e., the absence of trend observed while in the study is indicative of the missing data). Both examples are consistent with ignorable nonresponse.

If unmeasured characteristics of individuals or their treatment experience lead to dropout that affects the distribution of missing data, then nonresponse is nonignorable. For example, if a patient drops out of a study because of a side effect of the intervention, but side effects are not included as covariates in the model, the missing data would be nonignorable and the inferences drawn from these models would be invalid. This is not a consequence of using more complex statistical models. Use of sophisticated models helps explicate assumptions. On the other hand, simple models can lead to questionable conclusions. For example, the quasi-likelihood approach of Liang and Zeger (1986; Zeger & Liang, 1986) assumes no distributional form for the outcome measures and can therefore be applied to a wide variety of data (i.e., binary, ordinal, and continuous). The disadvantage however is that missing data are ignorable only if they are completely explained by the covariates in the model. Since no distributional form is assumed for outcomes, distribution of the missing data conditional on the observed outcomes is unknown and therefore cannot be used to justify statistical inferences in the presence of missing data. Indeed, in the presence of missing data, the quasi- or partial-likelihood approaches become even more restrictive than the full-likelihood procedure described here, in that the consistency of the quasi-likelihood estimates is now guaranteed only if the true correlation among repeated outcomes is known for each subject. This information, of course, is never available.

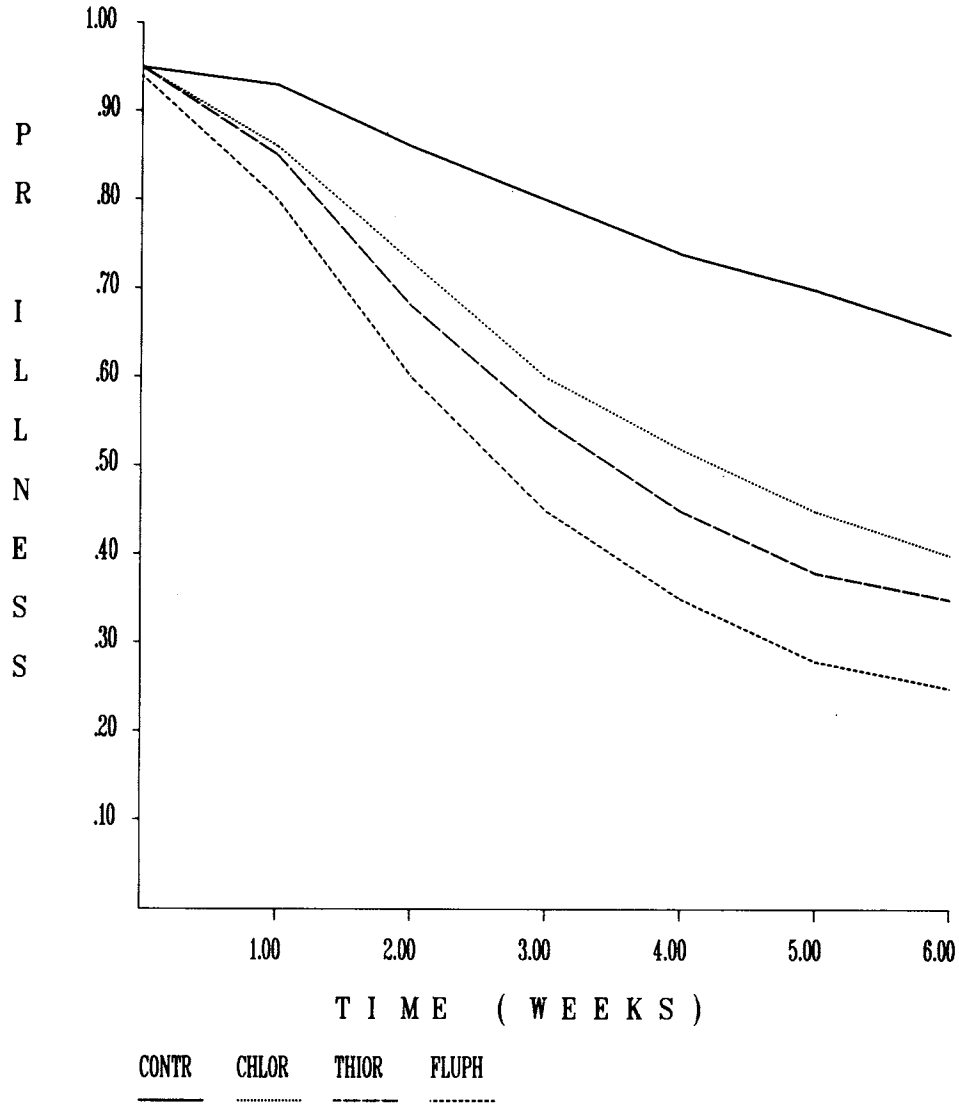

Figure 5. IMPS 79 Severity × Time interaction for men. IMPS 79 Severity = Item 79, "Severity of Illness," from the Inpatient Multidimensional Psychiatric Scale (Lorr & Klett, 1966). PR ILLNESS = probability of illness; CONTR = control; CHLOR = chlorpromazine; THIOR = thioridazine; FLUPH = fluphenazine.

Unfortunately, very little computer software is commercially available, and the models presented here are computationally heavy. A prototype computer program is available (MIXOR) from the National Institute of Mental Health Services Research Branch.

There are a number of directions for future research in this area. First, while the models presented here were developed for binary data, analyses can be devised for ordinal response data as well. The major difference is that the model now involves estimates of K-1 thresholds describing the point of transition from each response category to the next highest one in terms of underlying response strength (see Hedeker & Gibbons, 1994). Second, the model presented here assumes that conditional on the fixed and random effects included in the model the residual errors are independent and have constant variance. It is far more plausible that some degree of serial correlation among residual errors will be present, perhaps first-order autocorrelation. Stiratelli et al. (1984) suggest an approximate solution to this problem by including the observed outcome on the previous occasion as a covariate in the model. It is unclear whether this does yield independent residual errors or how these estimates are influenced by missing data. Gibbons and Bock (1987) suggest a direct approximation of the likelihood for the random-effects probit model that permits residual correlation and how the single parameter ρ of a first-order autoregressive error structure can be jointly estimated with the other fixed and random effects in the model. Unfortunately, error bounds for their approximation are still unknown, so their solution cannot be fully relied on at this time.

More recently, advances in Monte Carlo methods for numerical integration, for example, Gibbs sampling (Gelfand & Smith, 1990) have been developed with remarkable results.

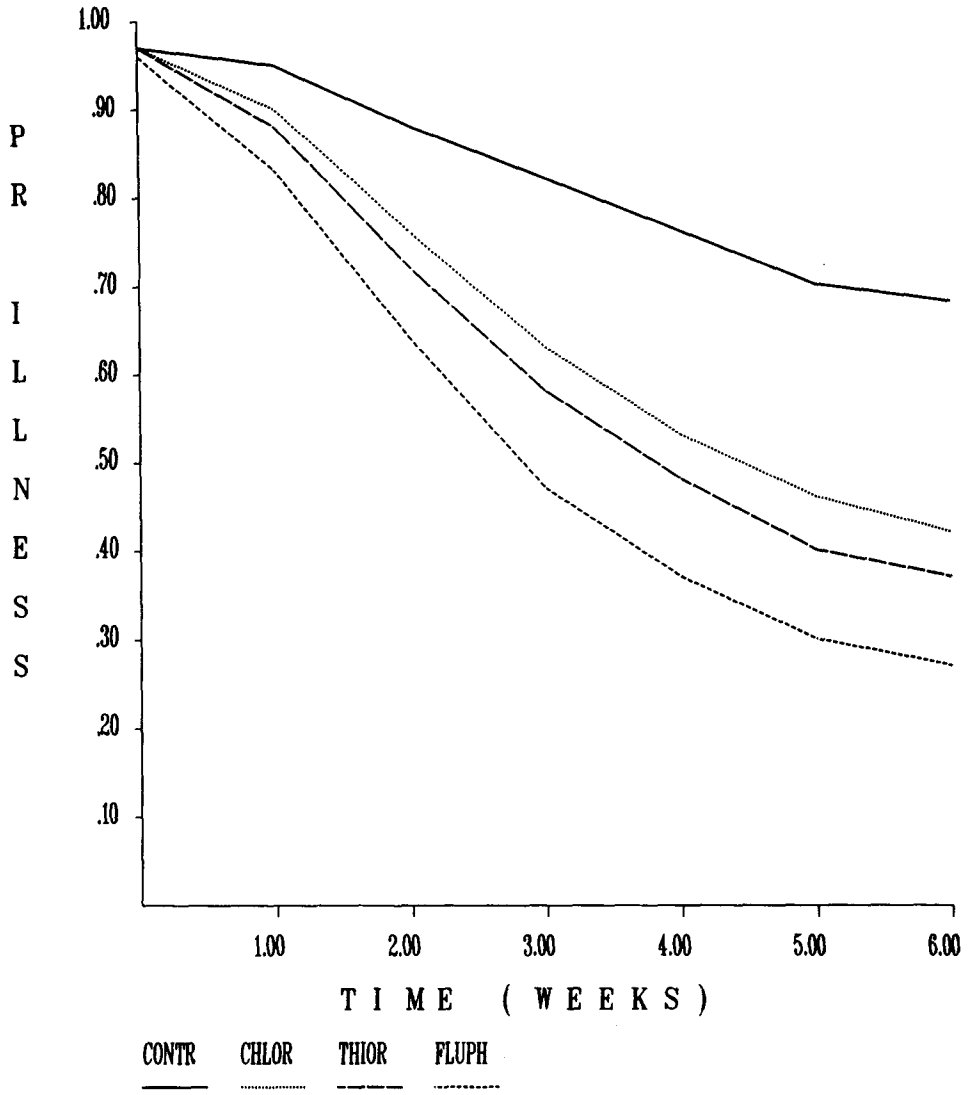

Figure 6. IMPS 79 Severity × Time interaction for women. IMPS 79 Severity = Item 79, "Severity of Illness," from the Inpatient Multidimensional Psychiatric Scale (Lorr & Klett, 1966). PR ILLNESS = probability of illness; CONTR = control; CHLOR = chlorpromazine; THIOR = thioridazine; FLUPH = fluphenazine.

These approaches can be readily adapted to the problem of evaluating the likelihood of correlated probit models as well.

It is often the case that data are both clustered and longitudinal. For example, in a multicenter clinical trial, subjects are nested within research centers and repeatedly measured over time. It may be reasonable to assume that the centers represent a random sample from a population of possible research sites and that observations within individuals and within centers will not be independent. Combining the two models presented here into a three-level model (i.e., center, subject, and measurement occasion) would have widespread application.

References

Anderson, D., & Aitkin, M. (1985). Variance components models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B, 47,* 203–210.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D. (Ed.). (1989). *Multilevel analysis of educational data.* San Diego, CA: Academic Press.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Conoway, M. R. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the American Statistical Association, 84,* 53–61.

Finney, D. J. (1971). *Probit analysis.* Cambridge, England: Cambridge University Press.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398–409.

Gibbons, R. D., & Bock, R. D. (1987). Trend in correlated proportions. *Psychometrika, 52,* 113–124.

Gibbons, R. D., Hedeker, D., Waternaux, C. M., & Davis, J. M. (1988). Random regression models: A comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin, 24,* 438–443.

Gibbons, R. D., Hedeker, D. R., Charles, S. C., & Frisch, P. (in press). A random-effects probit model for predicting medical malpractice claims. *Journal of the American Statistical Association.*

Gibbons, R. D., Hedeker, D. R., Elkin, I., Waternaux, C., Kraemer, H. C., Greenhouse, J. B., Shea, M. T., Imber, S. D., Sotsky, S. M., & Watkins, J. T. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Archives of General Psychiatry, 50,* 739–750.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Oxford University Press.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika, 78,* 45–51.

Hedeker, D. R., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics.*

Hedeker, D., Gibbons, R. D., Waternaux, C. M., & Davis, J. M. (1989). Investigating drug plasma levels and clinical response using random regression models. *Psychopharmacology Bulletin, 25,* 227–231.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 42,* 805–820.

Koch, G., Landis, J., Freeman, J., Freeman, H., & Lehnen, R. (1977). A general methodology for the analysis of experiments with repeated measurements of categorical data. *Biometrics, 33,* 133–158.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine, 7,* 305–315.

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

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13–22.

Lorr, M., & Klett, C. J. (1966). *Inpatient Multidimensional Psychiatric Scale: Manual,* (rev.). Palo Alto, CA: Consulting Psychologists Press.

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

Ware, J. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician, 39,* 95–101.

Waternaux, C. M., Laird, N. M., & Ware, J. H. (1989). Methods for analysis of longitudinal data: Blood lead concentrations and cognitive development. *Journal of the American Statistical Association, 84,* 33–41.

Willett, J. B., Ayoub, C. C., & Robinson, D. (1991). Using growth modeling to examine systematic differences in growth: An example of change in the functioning of families at risk of maladaptive parenting, child abuse, or neglect. *Journal of Consulting and Clinical Psychology, 59,* 38–47.

Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80,* 513–524.

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42,* 121–130.

Received September 30, 1991
Revision received July 13, 1993
Accepted July 19, 1993 ■