Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 5

11-1-2004

Modeling Incomplete Longitudinal Data

Hakan Demirtas
University of Illinois at Chicago, demirtas@uic.edu

DOI: 10.22237/jmasm/1099267500

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 305-321

1538 – 9472/04/$95.00

Regular Articles

Modeling Incomplete Longitudinal Data

Hakan Demirtas
School of Public Health
University of Illinois at Chicago

This article presents a review of popular parametric, semiparametric and ad-hoc approaches for analyzing incomplete longitudinal data.

Key words: longitudinal data, missing data, multiple imputation, ignorability, estimating equations, selection models, pattern-mixture models

Introduction

Missing observations are common in longitudinal studies. This article focuses on attrition, where responses are available for a subject until a certain occasion, and missing for all subsequent occasions. In the presence of incomplete data, the risk of reaching incorrect decisions is higher, because missing data may degrade the performance of confidence intervals, bias parameter estimates and reduce statistical power. Handling incomplete data generally requires special techniques and inferential tools. In this article, commonly used ad-hoc methods, semiparametric methods and likelihood-based models for incomplete repeated-measures data were reviewed and these approaches were applied to a real dataset.

The real data example pertains to a psychiatric trial in which dropout behavior appears to be quite different in the treatment and control groups. Data were obtained from the National Institute of Mental Health Schizophrenia Collaborative Study, where patients were randomly assigned to receive one of three anti-psychotic medications or a placebo.

Hakan Demirtas is an Assistant Professor of biostatistics at the University of Illinois at Chicago. His research interests are the analysis of incomplete longitudinal data, multiple imputation and Bayesian computing. E-mail address: demirtas@uic.edu

As noted by Hedeker and Gibbons (1997), performance of the three drugs was quite similar; following their approach, the subjects from the three drug treatments were collapsed into a single group. The outcome of interest, severity of illness, was measured on an ordinal scale ranging from 1 (normal) to 7 (extremely ill), which was treated as continuous. Measurements were planned for weeks 0, 1, 3, and 6, but missing values occurred primarily due to dropout. A few patients had missing measurements and subsequently returned; for simplicity these have been removed. A small number of measurements were also taken at intermediate time points (weeks 2, 4, and 5) which were also ignored. These exclusions reduced the sample from 1,603 subject-observations to 1,500.

With these exclusions, the sample contains 312 patients who received a drug and 101 who received a placebo. In the drug group, 3 patients dropped out immediately after week 0, 27 dropped out after week 1, 34 dropped out after week 3, and 248 completed the study. In the placebo group, no patients dropped out after week 0, 18 dropped out after week 1, 19 dropped out after week 3, and there were 64 completers. In this trial, the mean profile for placebo group is slightly declining, indicating mild improvement over time, but the drug group declines more dramatically. Dropout affects the two groups differently. If patients are classified as dropouts or completers, the dropouts in the placebo group appear to be more severely ill

than the completers and show less improvement. In the drug group, however, the opposite occurs: dropouts appear to be less severely ill than completers and improve more rapidly. Mean profiles for dropouts and completers in the two groups are shown in Figure 1. One plausible explanation is that those receiving the placebo who experience little or no improvement may be leaving the study to seek treatment elsewhere. On the other hand, those in the drug group who improve dramatically may be dropping out because they feel that treatment is no longer necessary.

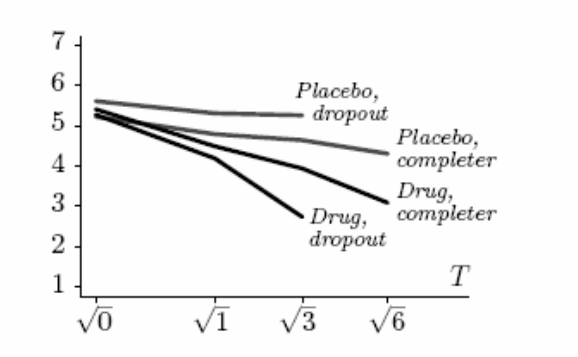

Figure 1. Mean observed response in psychiatric trial by treatment group (placebo, drug) and dropout status (dropout, completer), plotted versus *T* = square root of week.

Organization of this article is as follows: An overview is provided with background information on incomplete longitudinal data and ignorability. Popular longitudinal modeling techniques such as linear and nonlinear mixed models, semiparametric marginal approaches and their weighted versions, single imputation and its variants, multiple imputation, selection and pattern-mixture models are presented along with the implications of missing data for these commonly used methods. In the portion dealing with application, most of the mentioned methods were applied to the psychiatric trial dataset and findings were compared. Conclusions include remarks and discussion stressing the importance of sensitivity analyses and robustness studies.

Overview

Mechanisms for missing data and dropout

The properties of missing-data methods depend on the manner in which data became missing; every missing-data technique makes implicit or explicit assumptions about the missing-data mechanism. In this section, major classes of missing-data mechanisms were discussed, emphasizing the taxonomy introduced by Rubin (1976).

Many missing-data procedures in use today assume that missing values are missing at random (MAR) (Rubin, 1976). Let Y denote the complete set of responses for all subjects, and suppose that the distribution of Y depends on a set of unknown parameters of interest θ. Let R be the associated set of missing-value indicators. The elements of R take the values 1 or 0, indicating whether the corresponding elements of Y are observed or not. The conditional distribution of R given *Y* depends on the set of parameters δ. Let $Y = (Y_{obs}, Y_{mis})$ denote the partition of the data into the respective sets of observed and missing values. Finally, let (y_{obs}, r) be the realized value of (Y_{obs}, R). The missing values are said to be MAR if $P(R = r \mid Y_{obs} = y_{obs}, Y_{mis}, \delta) = P(R = r \mid Y_{obs} = y_{obs}; \delta)$ holds for all possible δ. Under MAR, the probability distribution of the indicators of missingness may depend on the observed data but must be functionally independent of the missing data. Intuitively speaking, MAR means that once appropriate account is taken of what have been observed, there remains no dependence of the missingness on unobserved quantities. A simple example is a two-occasion study of blood pressure where subjects are called back for the second measurement if the first measurement is high. This example is MAR because missingness on the second measurement depends only on the value of the first measurement which is always observed.

An important special case of MAR is missing completely at random (MCAR). Under MCAR, $P(R = r \mid Y_{obs} = y_{obs}, Y_{mis}; \delta) = P(R = r; \delta)$ for all possible δ. In this case, the response probabilities are independent of both the observed and unobserved parts of the dataset.

Suppose, for example, in a two-occasion study of blood pressure, a randomly chosen subset of subjects is called back for a second measurement. In this case, the missing-data mechanism is MCAR, because the probability that the second measurement is missing does not depend on blood pressure at either occasion.

If MAR is violated, the response probabilities depend on unobserved data; in this case, the missing values are said to be missing not at random (MNAR). MNAR situations require special care; to obtain correct inferences, one must specify a joint probability model for the complete data and the indicators of missingness.

Types of dropout

When missing data arise only through dropout, R can be summarized in a single variable that records the first time at which a value is missing or the time of a subject's last observed measurement. Special terminology has evolved for dropout, and this terminology is best understood by its relationship to MAR, MCAR and MNAR (Diggle & Kenward, 1994; Little 1995; Verbeke & Molenberghs, 2000).

Under MAR dropout, the probability of dropout may depend on observed covariates and past responses. Nonignorable dropout (ND) is used interchangeably with MNAR. Under ND, the dropout probability may depend on unobserved covariates, current and future unobserved responses. Little (1995) clarified the role of covariates in this classification scheme. He used "covariate-dependent dropout" (CDD) for the situation where dropout may depend on completely observed covariates. Under CDD,

$$P(R = r \mid Y_{obs} = y_{obs}, Y_{mis}, x; \delta) = P(R = r \mid x; \delta)$$

where x is the realized value of fully observed covariates X. A clinical trial where dropout rates differ among treatment groups, but otherwise unrelated to responses, would be an example of this type. Diggle and Kenward (1994) use the terms *random dropout* (RD) for MAR dropout, *informative dropout* (ID) for nonignorable dropout, and completely random dropout (CRD) if the dropout does not depend on responses or covariates. Little's terminology is more consistent with the literature on general missing-data problems, because the term completely random has historically been reserved for situations where missingness does not depend on any variables at all.

Ignorability

An important concept in the theory of missing data, closely related to MAR, is ignorability. A missing-data mechanism is ignorable if (a) the missing data are MAR and (b) the parameters δ and θ are distinct (Little & Rubin, 2002). From a frequentist perspective, distinctness means that the joint parameter space of (δ, θ) is the Cartesian cross-product of the individual parameter spaces for δ and θ. From a Bayesian perspective, it means that the joint prior distribution of (δ, θ) factors into independent priors for δ and θ (Schafer, 1997a).

The term ignorable suggests that the missing-data mechanism can, in some sense, be ignored when performing statistical analyses. Rubin (1976) precisely explained what it means to ignore the missing-data mechanism, both from frequentist and likelihood/Bayes standpoints, and provided conditions under which ignoring the missing-data mechanism is valid for inferences about θ. In the frequentist case, ignoring the missing-data mechanism means fixing R at its realized value and using $P(Y_{obs} \mid R = r; \theta, \delta)$ as a repeated-sampling distribution. That is, it is pretended that Y_{obs} is the data that had been intended to collect. In the likelihood/Bayes situation, ignoring the missing-data mechanism means using

$$\int P(Y_{obs} = y_{obs}, Y_{mis}; \theta) dY_{mis}$$

as the likelihood function for θ. The conditions under which these approaches are valid differ. In the likelihood/Bayes case, ignoring the missing-data mechanism is valid when are distinct and the missing data are MAR. In the frequentist case, the stronger condition of MCAR is needed.

This definition of ignorability seems to implicitly assume that one is working within a likelihood-based or Bayesian context. The reason why the missing-data mechanism can be

ignored under this condition is that joint log-likelihood for δ and θ partitions as $l(\theta,\delta;y_{obs},r)=l(\theta;y_{obs})+l(\delta;r)$. Information about the complete-data population parameter is contained fully in the first term; inferences about θ are unaffected by R, and there is no need to model $P(R=r\mid y,\delta)$. However, if one is not working in likelihood-based or Bayesian frameworks, one may need to formally model R even when the missing data are MAR. Therefore, the appropriateness of not modeling the missing-data process is not a property of the mechanism alone, but a property of the mechanism and the method of analysis.

The precise meaning of ignorability and its implications have often been misunderstood and misapplied, because many statistical procedures in use today are actually a hybrid of likelihood and frequentist approaches. For example, the use of an expected information matrix is frequentist, because it takes an expectation over the distribution of all possible data values. Helpful discussion and clarification of this point is given by Kenward and Molenberghs (1998).

Nonignorable modeling

Any violation of MAR leads to a nonignorable missingness mechanism. No simplification of the joint distribution is possible, and inferences can only be made about marginal responses by making further assumptions about which the observed data alone carry no information (Little & Rubin, 2002; Little, 1995). Under MNAR, the missingness mechanism does not drop out of the likelihood; the missingness indicators provide information about the parameters of the complete-data population. In these situations, assuming MAR may lead to biased estimates of parameters of the complete-data population; joint modeling of longitudinal response and dropout mechanism is needed.

Completers only analysis

Omitting the subjects with missing observations tends to introduce bias, to the extent that the incompletely observed cases differ systematically from the complete cases. Completers may be unrepresentative of the population for which the inference is usually intended: the population of all cases, rather than the population of cases with no missing data. In longitudinal studies with human or animal subjects, not all subjects complete the study and especially when completers and dropout seem to follow different trajectories, analyzing only the completers may be very misleading and inefficient.

Last observation carried forward (LOCF)

LOCF is often used in the analyses of clinical trials for FDA (Food and Drug Administration). It tends to understate differences in estimated time trends between treatment and control groups. Although LOCF is thought to be conservative, standard errors are biased downward as well, so it is not necessarily conservative. LOCF seems appealing only when between subject variation is high but responses within a subject is relatively stable over time. In this case, last observation may be a decent predictor for missing data points.

Mean imputation

Imputing the subject-mean seriously distorts trends over time and within-subject covariance structure. Imputing the occasion-mean distorts trends within subjects and between-subject variation. Both mean imputation methods introduce bias into longitudinal analyses and seriously impair standard errors and hypothesis tests.

Other single imputation techniques

Imputing from conditional means (e.g. through a regression prediction), from unconditional distributions (e.g. hot deck) or conditional distributions (through a predictive distribution) have been applied to longitudinal data, but the shortcomings of these methods have been well-documented (Little & Rubin, 2002; Schafer & Graham, 2002).

Single imputation strategies outlined above are designed to precisely predict the missing values. However, the goal of a missing-data procedure is to draw accurate inferences about the population quantities (e.g. mean change over time), not to accurately predict missing values. With imputation, the best way to achieve this goal is to preserve all aspects of the

data distribution (means, trends, within- and between-subject variation, etc.). Ad-hoc imputation methods inevitably preserve some aspect, but distort others.

Multiple imputation

Multiple imputation (MI) is a Monte Carlo technique (Rubin 1987, 1996) in which the missing values are replaced by a set of $m>1$ simulated versions of them. These simulated values are drawn from a Bayesian posterior predictive distribution for the missing values given the observed values and the dropout times.

Carrying out MI requires two sets of assumptions. First, one must propose a model for the data distribution which should be plausible and should bear some relationship to the type of analysis to be performed. In the case of longitudinal analyses, the model should be capable of preserving the correlation structure and time trends within individuals. The second set of assumptions pertains to type of missingness mechanism. An assumption of MAR is commonly employed for MI. However, the theory of MI does not necessarily require MAR; MI may also be performed under nonignorable models.

The key idea of MI is that it treats missing data as an explicit source of random variability to be averaged over. The process of creating imputations, analyzing the imputed datasets, and combining the results is a Monte Carlo version of averaging the statistical results over the predictive distribution of the missing data,

$$\int P(\theta \mid Y)P(Y_{mis} \mid Y_{obs})dY_{mis}.$$

In practice, a large number of multiple imputations is not required; sufficiently accurate results can often be obtained with $m \le 10$. Once the imputations have been created, the m completed datasets may be analyzed without regard for dropout; all relevant information on nonresponse is now carried in the imputed values. Once the quantities have been estimated, the m versions of the estimates and their standard errors are combined by simple arithmetic as described by Rubin (1987). Let $\hat{Q}^{(j)}$ and $\sqrt{U^{(j)}}$ denote the estimate and standard error for a scalar population quantity Q obtained from imputed dataset $j=1,...,m$. The overall estimate of Q is $\bar{Q}=m^{-1}\sum_j \hat{Q}^{(j)}$, and the overall standard error is $\sqrt{T}$, $T=\bar{U}+(1+m^{-1})B$, where $\bar{U}=m^{-1}\sum_j U^{(j)}$ and $B=(m-1)^{-1}\sum_j(\hat{Q}^{(j)}-\bar{Q})^2$. Interval estimates and tests may be based on the approximation $(\bar{Q}-Q)T^{-1/2}\sim t_\gamma$, where $\gamma=(m-1)(1+r^{-1})^2$, $r=(1+m^{-1})B/\bar{U}$, and the estimated rate of missing information is approximately $r/(1+r)$. Other rules for combining multidimensional estimates and test statistics are reviewed by Schafer (1997a Chap. 4).

MI may not be the best choice for every analysis, but it is a handy statistical tool and a valuable addition to a researcher's methodological toolkit. MI is attractive for a number of reasons. First, it allows researchers to use their favorite models and software; an imputed dataset can be analyzed by virtually any method that would be appropriate if the data were complete. Second, there are many classes of problems for which no direct ML procedure is available. For example, in longitudinal analyses, there is no direct ML method for incomplete covariates when occasions of measurement vary by individual. Third, MI singles out missing data as a source of random variation distinct from ordinary sampling variability. Finally, the separation of the imputation stage from the analysis stage provides flexibility to the entire modeling process.

Simple hypothesis testing and classical analysis of variance (ANOVA)

Let $y_i=(y_{i1},y_{i2,\ldots,}y_{ip})^T$ denote the responses for subject i, $i=1,2,...,m$ at a common sets of occasions $t=(t_1,t_2,...,t_p)$. If there are no missing values, it is said that the data are balanced in the sense that all subjects are measured at a common set of occasions. Simple t-tests based on change in scores (e.g.

$y_{ip} - y_{i1}$) can be used to test the mean equality hypothesis. As a generalization, one may assume $y_i \sim N(\mu, \Sigma)$ where $\mu = (\mu_1, \mu_2, ..., \mu_p)^T$. The classical ANOVA decomposition for repeated measures can be used to determine if means at each time point are equal. Let SS_A, SS_B and SS_{AB} denote sums of squares for subjects, occasions and subject-occasion interactions with degrees of freedom 1, p-1 and (m-1)(p-1), respectively. Under the null hypothesis that all occasion-means are equal ($\mu_1=\mu_2=...=\mu_p$), the test statistics $F = SS_A(m-1) / SS_{AB}$ is distributed as $F_{(p-1),(m-1)(p-1)}$ provided that Σ satisfies the Huynh-Feldt circularity condition ($Var(y_{ij} - y_{ij'}) = 2\lambda$ for $j \neq j'$, for some $\lambda>0$.) (Huynh & Feldt, 1970). One example of circularity is compound symmetry, which arises when $y_{ij} = \alpha_i + \mu_j + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim N(0, \sigma_\xi^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_\xi^2)$ so that $V(y_{ij}) = \sigma_\alpha^2 + \sigma_\xi^2$ and $Corr(y_{ij}, y_{ik}) = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\xi^2)$. When the circularity assumption is violated, one can use more general multivariate regression models in which Σ is allowed to be unstructured (Seber, 1984).

When missing values destroy the balance, data analysts sometimes discard the subjects until balance is restored, or they impute missing values in such a way that the sums of squares are not distorted so that procedures requiring balanced data may be applied (Dodge, 1985). In agricultural experiments or laboratory settings, data are often balanced or nearly so. But in longitudinal studies with human or animal subjects, measurements at common sets of occasions are unlikely, so classical ANOVA is less common in these situations.

Linear mixed models

Linear mixed models (Laird & Ware, 1982) extend classical ANOVA to handle unbalanced data by relying on improved computational methods. That is, the inferential strategy is changed from exact distributional results to ML estimation. In linear mixed models, the variation in subjects' longitudinal profiles arises at two levels: At the first level, the vector of repeated measurements for each subject is related to time and time-varying covariates by a relatively small number of estimated subject-specific regression coefficients.

At the second level, one relates these coefficients to additional time-varying and static covariates such as treatment, baseline characteristics, gender and so forth. The linear mixed-model paradigm combines these two stages into a single modeling procedure. These models—which are also known as multilevel models, linear mixed-effects models, random-effects models, random-coefficient models and hierarchical linear models—have been implemented in many software packages, including HLM (Bryk, Raudenbush & Congdon, 1996), MLwiN (Rasbash et al., 2000), the S-PLUS function lme (Pinheiro & Bates, 2000), SAS PROC MIXED (Littell et al., 1996) and Stata (Stata Corp., 1997).

Adopting the notation of Laird and Ware (1982), let $y_i = (y_{i1}, y_{i2, \ldots,} y_{in_i})^T$ denote the responses for subject i. The number of responses and the times of measurement may vary arbitrarily from one subject to another. The model is

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \qquad (1)$$

where X_i ($n_i \times p$) and $\mathbf{Z}_i$ ($n_i \times q$) contain covariates, β are fixed effects, and b_i and ε_i are unobservable random errors distributed as

$$b_i \sim N_q(0, \psi) \qquad (2)$$

$$\varepsilon_i \sim N_{n_i}(0, \sigma^2 V_i) \qquad (3)$$

independently for $i = 1, ..., m$. In this model, the vector of repeated measurements on each subject follows a linear regression model where some of the regression coefficients are common to population, whereas other coefficients vary by subject. Because the model does not assume any particular form for X_i and Z_i, it can handle time-varying covariates and unequally spaced

responses. The columns of Z_i usually span a subspace of the linear space spanned by X_i. Centering the distribution of b_i at zero causes β to become the population-averaged regression coefficients, and the random effects $b_1,...,b_m$ become perturbations due to inter-subject variation. When the number of measurements is small, the identity matrix (I_i) is typically used for V_i. Patterned correlation structures (auto-regressive, banded) are possible, in which case V_i contains some unknown parameters.

Averaging over the distribution of the latent random effects b_i, the marginal distribution of y_i is

$$y_i \sim N(X_i\beta, \Sigma_i), \qquad (4)$$

where $\Sigma_i = Z_i\psi Z_i^T + \sigma^2 V_i$. Therefore, the elements of β represent the effects of covariates in X_i on the mean response, both for a single subject (i.e. given b_i) and on average for the population.

When the data entering the linear mixed model are unbalanced by design, ML estimation using a likelihood derived from (4) is entirely appropriate. If some responses for some subjects are missing, one may omit the missed occasions and apply ML to the reduced data; this is appropriate if the missing responses are MAR.

Nonlinear mixed models

Nonlinear mixed models generalize the linear mixed models to situations where the response is not necessarily normal. They are also known as generalized linear mixed models or generalized linear models with random effects. In these models, one supposes that y_{ij} belongs to an exponential family with $E(y_{ij}) = \mu_{ij}$ and $\mu_i = (\mu_{i1},..., \mu_{in_i})^T$. The link function—the function that determines the relationship between expected mean and covariates— is $h(\mu_i) = X_i\beta + Z_i b_i$. If *h* is the identity function and the responses are normal, then this reduces to a linear mixed model. More generally, the nonlinear mixed model can be applied to repeated observations of binary and count variables.

Except in special cases, the likelihood function for nonlinear mixed models

$$L = \prod_i \int P(y_i \mid b_i) P(b_i) db_i \qquad (5)$$

cannot be computed analytically; it can only be approximated by numerical techniques such as Gauss-Hermite quadrature (Abramowitz & Stegun, 1964), adaptive quadrature (Kronrod, 1965) and Laplace expansions (Stroud, 1971). Algorithms for maximizing (5) are considerably more complicated than for the normal linear mixed model. Early programs used a technique called penalized quasi-likelihood (PQL) (Breslow & Clayton, 1993), whereas later programs (HLM, MLWin, PROC NLMIXED) use true ML. True ML is better, because the resulting estimates tend to be less biased. Bayesian inference is also possible by Markov Chain Monte Carlo (MCMC) (e.g., Spiegelhalter et al., 1999).

In the linear mixed model, β is the effect of X_i on μ_i both for a single subject and on average for the population. In the nonlinear case, however, the distinction between subject-specific (SS) and population- averaged (PA) effects naturally emerges: $E(y_{ij} \mid b_i) = h^{-1}(X_i\beta + Z_i b_i)$ is the SS mean response, whereas,

$\mu_{ij} = E[E(y_{ij} \mid b_i)] = \int h^{-1}(X_i\beta + Z_i b_i) dP(b_i)$

is the PA mean response. SS and PA effects have different interpretations and are appropriate in different circumstances (Zeger, Liang & Albert, 1988).

When missing data appear in nonlinear mixed models, as long as true ML or Bayesian techniques (not PQL) are used, the implications of missing responses are no different from normal linear mixed models; the procedures work as long as MAR is satisfied.

Semiparametric marginal models

Nonlinear mixed models are based on an SS formulation. Another way to formulate a model is to specify PA effects directly. Liang and Zeger (1986) proposed an estimation technique called generalized estimating equations (GEE) based on a multivariate version of quasi-likelihood (McCullagh & Nelder, 1989; Wedderburn, 1974). This formulation is semiparametric; rather than specifying a full distribution for the response, one only needs to specify its first two moments. That is, (a) the mean response as a function of covariates and (b) variances and covariances of the response as a function of the mean response are specified. In this approach, a broad class of non-Gaussian outcomes can be accommodated. Quasi-likelihood modeling is theoretically attractive, because it yields consistent and asymptotically normal estimates even when the covariance structure is misspecified. For this reason, GEE methodology has become quite popular for the analysis of longitudinal data.

The model is formulated as follows. Let $i = 1,...,m$ and n_i denote the subjects and the number of measurements for each subject, respectively. Let $\mu_i = (\mu_{i1},...,\mu_{in_i})^T$ be the expectation of $y_i = (y_{i1}, y_{i2,\ \ldots\ ,}\ y_{in_i})^T$ which is regarded as a function of covariates: $\mu_i = h^{-1}(X_i\beta)$ where β is a $p \times 1$ vector of unknown coefficients, X_i is an $n_i \times p$ covariate matrix, and *h* is the link function. The covariance matrix for y_i, denoted by V_i, is a function of μ_i (and hence β) and additional unknown parameters.

The estimate of β is obtained as the solution to the quasi-score equations

$$S(\beta) = \sum_{i=1}^{m} X_i^T \Delta_i A_i V_i^{-1}(y_i - \mu_i) = 0 \qquad (6)$$

where $\Delta_i = \partial\beta / \partial\mu_i$. The covariance matrix for y_i is usually parameterized as $V_i = A_i^{1/2} M_i(\alpha) A_i^{1/2} / \Phi$, where A_i is $n_i \times n_i$ an diagonal matrix with $g(\mu_{ij})$ as the j^{th} diagonal element; g is a hypothesized variance function; $M_i(\alpha)$ is a working correlation matrix and α is a vector that fully characterizes $M_i(\alpha)$; and Φ is a scale parameter. Therefore, the terms in equation (6) depend on β, α and Φ, but β is the parameter of interest whereas α and Φ are nuisance parameters. Solutions are obtained using iteratively reweighted least squares. At each iteration of the algorithm, one must plug in $\sqrt{m}-$consistent estimates of α and Φ; for details, see Liang and Zeger (1986).The solution to GEE, $\hat{\beta}$, is $\sqrt{m}-$consistent, asymptotically normal, and efficient if the hypothesized covariance structure is correct (Zeger and Liang, 1986). But the popularity of the method stems from the fact that approximate unbiasedness and normality hold even if assumptions about second moments are wrong (Diggle et al., 2003). If the assumed covariance structure is correct, a consistent estimator of *Cov*($\hat{\beta}$) is

$$(X^T \Delta A A^T \Delta^T X)^{-1} \qquad (7)$$

where X_i is the matrix of stacked X_i's, A is the stacked A_i's and Δ is the stacked Δ_i's. If $V_i \neq Cov(y_i)$, (7) can be be biased. In that case, however, a consistent estimator of *Cov*($\hat{\beta}$) can be obtained by the Huber-White information sandwich,

$$B\left[\sum_i X_i^T \Gamma_i (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)^T \Gamma_i^T X_i\right] B \qquad (8)$$

where $B = (X^T \Gamma \Gamma^T X)^{-1}$, $\Gamma = \Delta A$ and $\Gamma = \Delta_i A_i$. (Huber, 1967; White, 1980). In the literature (7) is often called a naive or model-based variance estimator, whereas (8) is called a robust or empirical variance estimator.

In practice, users of GEE typically select the variance function *g* based on the type of response variable. When y_{ij} is a frequency or count, for example, a natural choice is

$g(\mu_{ij}) = \mu_{ij}$. The working correlation matrix $M_i(\alpha)$ is chosen to reflect the hypothesized relationships among responses within subjects. Popular choices of $M_i(\alpha)$ include independence, exchangeable, one-dependent, auto-regressive or unstructured. In the independence model, $M_i(\alpha) = I$ and α is empty. Exchangeability means $M_i(\alpha) = (1-\alpha)I + \alpha 11^T$. In the one-dependent case, the $(t, t+1)^{th}$ element of M is taken to be α_t. Auto-regressive correlations can be expressed as $Corr(y_{ij}, y_{ik}) = \alpha^{|t_{ij} - t_{ik}|}$, where t_{ij} and t_{ik} are the observation times associated with y_{ij} and y_{ik}, respectively. Under the unstructured model, $M_i(\alpha)$ is completely unspecified. In that case, the data must be able to support the estimation of all unknown correlation parameters, which requires measurements at a relatively small number of common time points.

The GEE and sandwich methods attempt to "robustify" inferences by relaxing assumptions on the data model, but in doing so, they impose stronger assumptions on dropout mechanisms. The impact of missing data in GEE is quite different from parametric modeling. When elements of y_i are missing, one can omit the missed occasions for certain covariance structures. Liang and Zeger (1986) noted that if the working covariance assumptions are correct, the GEE estimator and the model-based covariance matrix (7) are consistent under MAR, because GEE then becomes maximum likelihood (ML). If the covariance assumptions are wrong, consistency of the GEE estimation and the information sandwich generally requires the missing data to be MCAR, because the sandwich has no likelihood interpretation. Work on weighted estimating equations (WEE) attempts to resolve this problem.

Joint models for longitudinal response and dropout

In practice, the hypothesis of random dropout is essentially untestable; it cannot be verified nor contradicted by examination of the observed data (Little & Rubin, 2002 Chap. 15). If this assumption is doubtful, alternative procedures should be developed, especially when the degree of departure from MAR is thought to be severe. When nonignorable missingness is suspected, it is necessary to make strong assumptions about the missingness mechanism and propose a specific model for it. That is, one needs to model the joint distribution of the longitudinal response and the dropout. From the likelihood point of view, there are two major ways to construct these models based on different factorizations of the joint distribution: selection models and pattern-mixture models.

Selection models

Selection models, which first appeared in the econometrics literature (Heckman, 1976; Amemiya, 1984), combine a model for the distribution of the complete data with a conditional model for the indicators of missingness given the data. In selection models (suppressing the parameters in the notation), the joint distribution of $f(y_i, r_i \mid x_i)$ is factored as $f(y_i \mid x_i) f(r_i \mid y_i, x_i)$. For example, one could assume that (a) a response variable follows a classical linear regression given a set of covariates, and (b) the probability that a response is observed is related to covariates and the response itself through a logit or probit regression function. These regression-type selection models have become a standard tool of econometricians (Maddala, 1983; Greene, 2000). The OSWALD software package (Smith et al., 1996) provides model-fitting routines for longitudinal data; this software is based on an extension of the work in Diggle and Kenward (1994).

Considering the responses and covariates to be the reasons for missingness, as a selection model does, can be intuitively appealing. Despite their conceptual appeal, the reputation of these models among statisticians is highly controversial. For example, Little and Rubin (2002, Chap. 15) argued that results from these models tend to be highly sensitive to departures from the assumptions about the shape of the complete-data population. In one example, Kenward (1998) demonstrated that a

slight perturbation to the population model—assuming a Student's t-distribution rather than a normal—caused drastic changes in parameter estimates. For these reasons, many statisticians tend to regard them as non-robust (see the discussion following the article by Diggle & Kenward, 1994).

Pattern-mixture models

Pattern-mixture models, a term coined by Little (1993), refers to the alternative strategy of first modeling the marginal distribution of the missingness indicators, and then the conditional distribution of the complete data given the pattern of missingness. The population of the complete data then becomes a mixture of distributions, weighted by the probabilities of the missingness patterns. Again, suppressing the parameters in the notation, $f(y_i, r_i \mid x_i)$ is factored as $f(r_i \mid x_i) f(y_i \mid r_i, x_i)$.

For example, consider a bivariate sample in which Y_1 is observed for all subjects but Y_2 is missing for some. A simple pattern-mixture model posits a Bernoulli distribution for R, a bivariate normal distribution for (Y_1, Y_2) given that $R = 1$, and another bivariate normal distribution for (Y_1, Y_2) given that $R = 0$. Because the conditional distribution of Y_2 given Y_1 is unobservable when $R = 0$, unverifiable assumptions must be made about this distribution in order to estimate aspects of the distribution of Y_2 in the full population. The assumptions of pattern-mixture models are no less strong than those of selection models, but some consider them to be more honest, because one knows precisely which parameters in the model formulation cannot be estimated from the observed data. Results from fitting these pattern-specific models are then averaged to obtain parameter estimates for the overall population (e.g. Hedeker & Gibbons, 1997). Alternatively, this process of averaging can be performed through multiple imputation (Glynn, Laird & Rubin, 1993).

Little (1995) defined two types of pattern-mixture models for nonignorable dropout: those with outcome-dependent dropout and those with random-effect-dependent dropout. In outcome-dependent models, subjects are grouped according to their dropout times and identifying restrictions are placed on the missing-value distributions for those groups (Little, 1993; Little & Wang, 1996; Molenberghs et al., 1998). In random-effect-dependent models, a random-coefficient model (1) is formulated with summaries of dropout time included as subject-level covariates (Wu & Bailey, 1989; Hedeker & Gibbons, 1997; Fitzmaurice et al., 2001). Little (1995) suggested that outcome-dependent models are appropriate when reasons for dropout seem closely related to the response variable itself, whereas random-effect-dependent models ascribe dropout to an underlying process (e.g. progression of a disease) which the outcome variable measures only imperfectly.

Weighted estimating equations

GEE may produce biased estimates if there are missing data, unless the data are MCAR. The method breaks down if the data are missing in a non-MCAR fashion, because the estimating equations on which they are based no longer have zero expectation. This problem suggests a method of modifying the estimating equations by applying weights which are proportional to the inverse-probabilities of response. Weighted estimating equations (WEE) that allow for non-MCAR missingness were first proposed by Robins, Rotnitzky and Zhao (1994, 1995). WEE are the semiparametric counterpart of joint modeling.

The price to be paid for incorporating weights is that a model must be specified for the missingness mechanism. Depending on the form of missingness model, WEE can handle MAR and MNAR mechanisms, but the parameters of an MNAR model are harder to estimate. Let W_i be an $n_i \times n_i$ matrix that contains the weights for subject i. W_i replaces the term $\Delta_i A_i V_i^{-1}$ in (6). So the information contained in Δ_i, A_i and V_i^{-1} about β and α is transferred to W_i. The weighted version of estimating equations becomes

$$S(\beta) = \sum_{i=1}^{m} X_i^T W_i (y_i - \mu_i) = 0 \qquad (9)$$

The weight matrix W_i is, in most cases, an $n_i \times n_i$ matrix whose j^{th} diagonal element is an estimate of the reciprocal-probability that the j^{th} element of y_i is observed. In that case, it is easy to see how the weighting scheme leads to a set of unbiased estimating equations. Modifying the notation a bit, let y_{ij}, μ_{ij}, w_{ij} and r_{ij} be the observed response, expected response, weight and missingness indicator, respectively for subject *i* at occasion *j*, respectively. The estimating equations become

$$S^w(\beta_k) = \sum_i \sum_j w_{ij} S_{ij}(\beta_k)$$

$$= \sum_i \sum_j w_{ij} x_{ijk} (y_{ij} - \mu_{ij}) = 0 .$$

$$w_{ij} = 1/P(r_{ij} = 1) = 1/E(r_{ij})$$

implies

$$E_R \left[E_y (S(\beta_k)) \right] = 0 .$$

(Carlin et al., 1999). In practice, the selection probabilities w_{ij}^{-1} are unknown and can, at best, be estimated by a logistic regression on similar-type of model for the r_{ij}'s. As shown by Robins et al. (1994, 1995), the asymptotic properties of the method are preserved if the inverse-weights w_{ij}^{-1} are $\sqrt{m}-$consistent estimates of the actual response probabilities.

In WEE, one is simply discarding the subject-occasions that are difficult to use because of missing responses and/or covariates, and reweighting the rest to make them seem more representative of the population. Robins, Rotnitzky and Zhao (1994) discard subject-observations with missing covariates. Robins, Rotnitzky and Zhao (1995) discard subject-observations with missing responses. Rotnitzky and Robins (1997), Rotnitzky, Robins and Scharfstein (1998) and Scharfstein, Rotnitzky and Robins (1999) discard various sets of subject-occasions for which covariates and/or responses are missing. The same idea is being applied in every case: estimating the inverse response probabilities using any information that seems to be related to missingness, including static covariates, time-varying covariates, baseline measures, pre-dropout responses or even post-dropout responses. With a post-dropout response, however, the influence on the response probability can only be guessed.

Application

Regarding the psychiatric dataset that was introduced before, Hedeker and Gibbons (1997) noted that the mean response profiles are approximately linear when plotted against the square root of week, and they express time on the square-root scale in their models. Adopting this convention, *T* (time) is defined to be the square root of week, and the time of last measurement *R* (which will be relevant in pattern-mixture models) is also expressed on the square-root scale. Furthermore, let *G* be an indicator for treatment group (0=placebo, 1=drug) and *D* an indicator of dropout status (0=completer, 1=dropout). The treatment effect is defined to be the difference in average slopes between the drug and placebo groups. In other words, the parameter of interest is the treatment by time interaction (drug effect over time) *G*×*T*.

Two ad-hoc approaches (LOCF and completers only), model-based parametric approaches (selection and pattern-mixture models) and model-based semiparametric methods (unweighted and weighted generalized estimating equations) have been applied to this particular dataset and an estimate of treatment by time interaction and its standard error is obtained for each analysis method.

Model fitting procedures for selection models are implemented through OSWALD (Smith et al., 1996). It finds the most likely values of the data and dropout model parameters jointly by the simplex algorithm developed by Nelder and Mead (1965). It allows three components of variance: a random intercept between subjects (with variance υ^2), a measurement error realized independently between two responses (with variance τ^2) and a serial association component (with variance σ^2 and autocorrelation function

$\rho(u) = \exp(-\phi|u|)$. The marginal covariance matrix for y_i is $\sigma^2 H_i + \tau^2 I + \upsilon^2 J$, where $H_i = \rho(|t_{ij} - t_{ik}|)$, J is the matrix of ones and I is the identity. In linear mixed model notation it is equivalent to $Z_i \psi Z_i^T + \sigma^2 I$. Regression parameters for the data model part are interpreted in the same way as in linear mixed models. It is again assumed that $y_i = X_i\beta + Z_i b_i + \varepsilon_i$ where the columns of X_i are a constant (one), *G*, *T*, and *G*×*T*. The columns of Z_i are a constant and *T*. The dropout (*D*) is assumed to depend on the time of the measurement (*T*), the treatment group (*G*) and some function of responses (see below) through a logit link.

Pattern-mixture models are implemented by incorporating summaries of *R* and their interactions with *G* and *T* into the fixed effects design matrix (X_i) in Equation (1). Then, it is proceeded by multiple imputation (MI) to obtain simulated values that are drawn from a Bayesian posterior predictive distribution for the missing values given the observed values and the dropout times. To create MI's for missing elements of y_i in a random-coefficient model, first a prior distribution for β and the covariance parameters in ψ, σ^2 and V_i must be specified. Then, a random value of these parameters is drawn from their joint posterior distribution given the observed elements of y_i. Finally, the missing elements of y_i are drawn from their conditional distribution given the observed elements derived from the marginal model $y_i \sim N(X_i\beta, Z_i\psi Z_i^T + \sigma^2 V_i)$, with β, ψ, σ^2 and V_i replaced by their simulated values. Repeating these steps *m* times produces *m* multiple imputations of the missing responses. Applications of MI to pattern-mixture models have been described by Verbeke and Molenberghs (2000) and Thijs et al. (2002). MI without large-sample approximations is possible by Markov chain Monte Carlo (MCMC), as described by Liu et al. (2000). SAS PROC MIXED provides an MCMC procedure for simulating posterior draws of model parameters without large-sample approximations.

The PAN library for S-PLUS developed by Schafer (1997b) performs these computations rather quickly under conjugate priors for σ^2 and ψ (scaled inverted chi-square and inverted Wishart, respectively) and $V_i = I$. Important issues in using these techniques, including the choice of prior hyperparameters and monitoring convergence of the MCMC algorithm, are discussed in Schafer (2001). Once the imputations have been created, completed datasets are analyzed with a direct maximum likelihood approach under linear mixed effects model that includes *G*, *T* and *G*×*T*. Finally estimates from *m*=10 imputations are combined by Rubin's (1987) rules. For a deeper discussion of these issues, see Demirtas and Schafer (2003).

Estimating equations-based approaches (GEE and WEE) are implemented through the software package YAGS (yet another GEE solver). An intercept, *G*, *T* and *G*×*T* are included in the model. In the unweighted version (GEE), correlation structure has chosen to be "independence" and "exchangeable". In WEE, weights are estimated based on the inverse probability of being observed for every subject-occasion in the dataset. Two ignorable mechanisms were assumed where weights are estimated by a logistic regression in which outcome variable is response/nonresponse indicator and covariates are *T*, *G* and some function of responses (see below).

In what follows, SM stands for selection model, PMM stands for pattern-mixture model; GEE and WEE are as defined before. Other details are described below:

LOCF: The last available measurement is carried forward to fill in unobserved cells.

COMP-ONLY: Only subjects having full set of measurements are considered for the analysis.

SM-1: *D* depends on *G*, *T* and the previous response; assumes ignorability.

SM-2: Same as SM-1 except that *D* depends on the average of available responses rather than the previous response.

SM-3: Same as SM-1 and SM-2 except that *D* depends on the current response rather than previous responses; assumes nonignorable dropout.

SM-4: Same as SM-3 except that *D* depends on the current and previous response.

PMM-1 Pattern-mixture model with *T*, *G*, *D*, *G*×*T*, *D*×*T*, *G*×*D* and an intercept in the fixed effects part; random intercept and slope in the random part of the linear mixed model (1).

PMM-2: Same as PMM-1, except that a linear term is used for the time of last measurement (*R*) rather than *D*.

PMM-3: PMM that does the extrapolation within each pattern without borrowing any information from other patterns.

PMM-4: PMM that borrows information from completers for inestimable parameters.

PMM-5: Same as PMM-4 except that information is borrowed from the neighboring pattern rather than completers.

PMM-6: Same as PMM-5 except that information is borrowed from all available patterns (by a weighted average of estimable parameters from all other patterns) rather than the neighboring pattern.

GEE-1: Unweighted GEE with "independence" correlation structure.

GEE-2: Same as GEE-2 with "exchangeable" correlations.

WEE-1: Weighted version of GEE-1 where weights are assumed to depend on *T*, *G* and the average of observed responses for each subject.

WEE-2: Same as WEE-1 except that the previous response is used rather than the average of observed responses in weight calculations.

Conclusion

Estimated coefficients for drug effect over time (*G*×*T*) and their standard errors under different analysis methods are tabulated in Table 1. Estimated coefficients are varying in a fairly wide range as well as their standard errors. Although one can safely conclude that there is a drug effect over time, the true magnitude of this effect is disputable. True data model and dropout mechanism are rarely known in practice, therefore it is advisable that statisticians should attack the problem with the help of applied researchers/scientists to be more competent with discipline-specific issues. Subject-matter considerations are as important as the actual analysis method.

Another important issue is sensitivity. Models for incomplete data can be sensitive to untestable assumptions and/or inestimable parameters. Sensitivity analyses are universally acknowledged as crucial, because observed data cannot reveal the true missing-data mechanism. These analyses are usually conducted by applying a variety of models to one dataset to see how the estimated effects vary due to differing modeling assumptions. If our basic conclusions about effects of interest do not change drastically over this family, then the scientific validity of these conclusions is enhanced. Conversely, if the answers do exhibit great variation, drawing firm conclusions seems unwise. For examples of sensitivity analyses, see Little and Wang (1996) and Chapter 20 of Verbeke and Molenberghs (2000).

Robustness studies are less common than sensitivity analyses mentioned, but they are also extremely valuable. A robust method will perform well when applied to a variety of situations when its assumptions are not met. Considerations of robustness may allow us to prefer one model, $Model_1$, to another, $Model_2$, even when $Model_1$ and $Model_2$ achieve the same likelihood for the current data set. That is, if a variety of plausible joint population models is devised for response and dropout—different in nature but all tending to produce samples that resemble the observed data— and if, by simulation, it is discovered that $Model_1$ performs better than $Model_2$ across many of these populations, then there may be more of am inclination to trust $Model_1$ than $Model_2$.

Applying models to a variety of populations consistent with observed data is a useful tool to assess robustness of the models under consideration. These simulations can help us to answer important questions that are being

raised by potential users of nonignorable methods. When nonignorable dropout cannot be ruled out, robustness analyses are preferable to placing total faith in a single model. Although the truth is never known, a model that performs well under differing assumptions that yields simulated datasets which mimic the real data can be regarded as more trustworthy.

Although analyzing a real dataset using the proposed methodology is useful and insightful, simulations are needed to assess how well the method performs. Because there is no consensus among statisticians about which competing method is best, many advocate sensitivity analysis by trying a variety of method and then seeing what happens, and/or identifying parameters that are nearly or truly inestimable and varying them over a plausible range.

This approach is certainly valuable, but limited. Methods that fit the data equally well may give different estimates and intervals for parameters of interest. But, that does not mean that the methods are equally robust to departures from the assumed model. Another approach to sensitivity analysis is to simulate the performance of a method when its assumptions are wrong by proposing a variety of populations and dropout mechanisms capable of producing data like actually seen; then simulating behavior of various methods over repeated samples from each population; and identifying methods that seem to perform well for a variety of populations. Simulations driven by the latter approach are recommended to find arguably the best method that leads to accurate estimates and narrow, calibrated intervals under plausible population/dropout mechanisms.

Table 1: Estimated treatment effect with standard error.

Method	*Estimate*	*Standard Error*
LOCF	-0.61	0.11
COMP-ONLY	-0.36	0.08
SM-1	-0.74	0.13
SM-2	-0.69	0.12
SM-3	-0.81	-
SM-4	-0.77	-
PMM-1	-0.73	0.09
PMM-2	-0.75	0.09
PMM-3	-0.78	0.12
PMM-4	-0.99	0.15
PMM-5	-1.22	0.36
PMM-6	-0.95	0.17
GEE-1	-0.63	0.09
GEE-2	-0.66	0.10
WEE-1	-0.62	0.11
WEE-2	-0.69	0.11

References

Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions*. National Bureau of Standards Applied Mathematics Series, Number 55. U.S. Government Printing Office: Washington.

Amemiya, T. (1984). Tobit models: a survey. *Journal of Econometrics*, *24*, 3-61.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88,* 9-25.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *Hierarchical linear and nonlinear modeling with HLM/2L and HLM/3L programs*. Chicago: Scientific Software International, Inc.

Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for nonignorable dropout. *Statistics in Medicine*, *22*, 2553-2575.

Diggle, P. J., Heagerty, P., Liang, K., & Zeger, S. L. (2003). *Analysis of longitudinal data* (Second Edition). Oxford: University Press.

Diggle, P. J., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics, 43*, 49-94.

Dodge, Y. (1985). *Analysis of experiments with missing data*. New York: Wiley and Sons.

Fitzmaurice G. M., Laird, N. M., & Shneyer L. (2001). An alternative parameterization of the general linear mixture model for longitudinal data with nonignorable dropouts. *Statistics in Medicine*, *20*, 1009-1021.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association, 88*, 984-993.

Greene, W. H. (2000). *Econometric analysis* (Fourth Edition). Upper Saddle River, NJ: Prentice-Hall.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement, 5*, 475-492.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64-78.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Fifth Berkeley Symposium in Mathematical Statistics and Probability*, Berkeley: University of California Press, 221-133.

Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-istributions. *Journal of the American Statistical Association, 65*, 1582-1589.

Kenward, M. G. (1998). Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine, 17*, 2723-2732.

Kenward, M. G., & Molenberghs G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science, 13*, 236-247.

Kronrod, M. (1965). *Nodes and weights of quadrature formulas*. Consultants Bureau: New York. [English translation of Kronrod (1964)].

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

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

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Second Edition). New York: Wiley.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *84*, 125-134.

Little, R. J. A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121.

Little, R. J. A., & Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, *52*, 98-111.

Liu M., Taylor J. M. G., & Belin T. R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics,56*, 1157-1163.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Second Edition). London: Chapman & Hall.

Molenberghs, G. M., Michiels, B., Kenward, M. G., & Diggle, P. J. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica, 52*, 153–161.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308-313.

Pinheiro J. C., & Bates D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Rasbash, J., Browne, W., Goldstein, H., & Yang M. (2000). *A user's guide to MLwiN* (Second Edition). London: Institute of Education.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*, 846-866.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90*, 106-121.

Rotnitzky, A., & Robins, J. M. (1997). Analysis of semi-parametric regression models with nonignorable nonresponse. *Statistics in Medicine, 16*, 81-102.

Rotnitzky, A., Robins, J. M., & Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association, 93*, 1321-1339.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association, 91*, 473-520.

Schafer, J. L. (1997a). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. (1997b). *PAN: Multiple imputation for multivariate panel data, software library for S-PLUS*. University Park, PA: The Pennsylvania State University, Department of Statistics.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods, 7*, 147–177.

Schafer, J. L. (2001). Multiple Imputation with PAN. In A. G. Sayer and L.M. Collins (Eds.) *New methods for the analysis of change* 355-377. Washington, DC: American Psychological Association.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association, 94*, 1096-1146.

Seber, G. A. F. (1984). *Multivariate observations*. New York: Wiley.

Smith, D. M., Robertson, W. H., & Diggle, P. J. (1996). *Oswald: Object-oriented software for the analysis of longitudinal data in S*. Technical Report MA 96/192, Department of Mathematics and Statistics, University of Lancaster, United Kingdom.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1999). *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit.

Stata Corporation. (1997). *Stata reference manual*. College Station, TX: Stata Press.

Stroud, A. H. (1971). *Approximate calculation of multiple integrals*. Prentice Hall: Eaglewood Cliffs.

Thijs H., Molenberghs G., Michiels B., Verbeke G., & Curran D. (2002). Strategies to fit pattern-mixture models. *Biostatistics, 3*, 245–265.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika, 61*, 439-447.

White, H. (1980). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–25.

Wu, M. C., & Bailey K. R. (1989). Estimation and comparisons of changes in the presence of informative right censoring. *Biometrics*, *45*, 939-955.

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

Zeger, S. L., Liang, K., & Albert, S. A. (1988). Methods for longitudinal data: A generalized estimating equation approach. *Biometrics*, *44*, 1049–1060.