Methods for Analysis of Longitudinal Data with Nonignorable Drop-out

Abstract Longitudinal studies follow subjects over time to determine trends in their responses. When a subject drops out of the study, there is the issue of how to deal with their partial information. If drop-out is related to unobserved characteristics of the subject, then this is called nonignorable or informative missingness. Methods that don't adjust for this missing data mechanism can produce biased results, but the assumptions made about the mechanism can't be verified from the observable data. For this reason, sensitivity analyses are advisable. Selection models, mixture models, and semiparametric models have all been proposed to handle nonignorable drop-out, but often come with concerns about model identifiability. This paper reviews the methods developed for longitudinal data with nonignorable drop-out and discusses issues of model identifiability and sensitivity analyses.

1 Introduction

In longitudinal studies where subjects are scheduled to come in for regular follow-up appointments, drop-out is not an uncommon occurrence. Drop-out is a monotone pattern of missingness, meaning that no responses are observed after the subject has missed one follow-up. Rubin (1976) formalizes definitions of different mechanisms for missing data. If missingness is unrelated to the observed or unobserved data, then the mechanism is missing completely at random (MCAR). If missingness depends only on observed covariates and/or responses, then the data is missing at random (MAR).

This paper examines the case when drop-out depends on unobserved measurements. This is called informative or nonignorable drop-out (Diggle and Kenward 1994, Little and Rubin 2002). Little (1995) further makes the distinction between drop-out based on the unobserved response and drop-out based on some "latent variable that the outcome variable is measuring *with error*." The latter case is called random-coefficient dependent drop-out.

For longitudinal data with no missing observations, it is standard to use generalized linear mixed models (GLMM) or generalized estimating equations (GEE). The GLMM model is $h^{-1}\{E(\mathbf{Y}_i \mid \mathbf{b}_i, \mathbf{X}_i, \mathbf{Z}_i)\} = \mathbf{X}_i'\beta + \mathbf{Z}_i'\mathbf{b}_i$, where h is a link function and $\mathbf{b}_i$ are multivariate normal coefficients of the random effects. GEE solve $\sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta'} \mathbf{V}_i^{-1} [\mathbf{Y}_i - \mu_i(\beta)] = 0$, where $\mathbf{V}_i$ is the covariance matrix of the response vector, $\mathbf{Y}_i$ (Liang and Zeger 1986). When data is MCAR, these methods can still be done on the complete observations since completers can be considered a random sample of all of the subjects. However, this will decrease the power of our analysis due to the smaller sample size (Davidian 2005).

When the data is MAR or nonignorably missing, the completers represent a biased sample of all of the subjects. Laird (1988) states that the non-response model must be taken into account when performing inference or the results could be biased. For example, if sicker subjects tend to drop out more often, then the results could be biased towards the response of the healthy subjects, thus not capturing the true effect.

Imputation methods and GEE methods with inverse weights to adjust for the probability of drop-out are valid methods for dealing with MAR and MCAR which use all available data. Imputation methods aim to "fill in" the missing data with estimates, and then proceed with the standard analysis. Imputed values are usually generated from the conditional distribution of the missing data given the observed data, $f(\mathbf{Y}_i^m \mid \mathbf{Y}_i^o, \mathbf{X}_i)$. The EM algorithm used with standard likelihoods imputes missing values with their conditional means given the observed data and the previous iteration's parameter estimates as part of the "E" step (Fitzmaurice 2011). It is clear that both methods take the MAR assumption into account, but this is not sufficient for nonignorable missingness.

It is important to note that it is possible to test whether the data is MCAR versus MAR or nonignorably missing, but it is not possible to formally test whether the data is MAR versus nonignorably missing. Therefore, the assumption about the missing data mechanism can have a huge effect on the analysis and needs to be explored with sensitivity analyses (Molenberghs 2009).

There are three major classes of models for dealing with nonignorable drop-out: Selection models, mixture models, and semiparametric models. The first two put parametric models on both the data structure and the missing data mechanism. When the number of parameters to estimate in these models gets large, this causes complex model identifiability issues. Additional restrictions and assumptions are necessary to alleviate this issue. Semiparametric methods aim to have less restrictive models by focusing on a broader class of models, but often unverifiable model assumptions are still necessary for identifiability. For this reason, many authors advocate approaching nonignorable missing data in a sensitivity framework.

This paper will review the developments in analyzing longitudinal data with nonignorable drop-out. Section 2 will present notation and some key assumptions. Section 3 reviews likelihood methods including selection and mixture models, while section 4 discusses identifiability. Section 5 introduces semiparametric methods, and section 6 will study a few different approaches to performing a sensitivity analysis. Finally, section 7 will provide a brief conclusion.

2 Notation and assumptions

Similar to the notation in Fitzmaurice and Laird (2000) and Little (2009), let $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})'$ be the complete response vector for subjects $i = 1, ..., N$, measured at time points $\{t_{i1}, ..., t_{in_i}\}$. It is assumed that the subjects are independent of one another. Often patients are scheduled to be measured an equal number of times so $n_i = n$. The response vector can be partitioned into the observed and missing parts, $\mathbf{Y}_i = (\mathbf{Y}_i^{o'}, \mathbf{Y}_i^{m'})'$. The $p \times 1$ vector $\mathbf{X}_{ij}$ holds the covariates of subject i at time t_j, and $\mathbf{X}_i = (\mathbf{X}_{i1}, ..., \mathbf{X}_{in_i})'$ is the $n_i \times p$ matrix of complete covariates for subject i. Let $\bar{\mathbf{X}}_{ij} = (\mathbf{X}_{i1}, ..., \mathbf{X}_{ij})$ be the history of covariates up through time t_j, and similarly let $\bar{\mathbf{Y}}_{ij} = (Y_{i1}, ..., Y_{ij})$ be the history of responses through time t_j.

Take D_i to be the drop-out time for subject i. If $D_i = t_j \leq t_{n_i}$ then the subject has dropped out between times t_{j-1} and t_j. If $D_i > t_{n_i}$ then the subject is a completer which will be notated

as $D_i = t_{n_i+1}$. Drop-out time is usually assumed to be discrete, but could be considered continuous for some semiparametric methods. Let $\mathbf{R_i} = (R_{i1}, ..., R_{in_i})'$ be the indicator vector of observed responses, where $R_{ij} = 1$ means the subject i has been observed at time t_j and $R_{ij} = 0$ otherwise. It is assumed that every subject has an observed response at the first time point, so $D_i > t_1$ and $R_{i1} = 1$. The observed data is $\{D_i, \mathbf{R_i}, \mathbf{Y}_i^o, \mathbf{X}_i\}$ for $i = 1, ..., N$. Some methods may accommodate missing covariates as well.

Unless otherwise stated, monotone missingness is assumed. This means that if $R_{ik} = 0$, then $R_{ij} = 0$ for all $j > k$. Once the subject fails to be observed at one time point, then there are no observed responses for them for the subsequent time points. This missingness pattern is consistent with a subject dropping out of a study, although the issue of drop-out along with intermittent missingness is also of interest.

3 Likelihood Methods

Inference about parameters that form the complete data likelihood, $L(\gamma \mid \mathbf{Y}, \mathbf{X}) = c\sum_{i=1}^{N} f(\mathbf{Y_i} \mid \mathbf{X}_i, \gamma)$, or conditional means of $\mathbf{Y} \mid \mathbf{X}$ are of interest. There are two major classes of parametric models that are used for data with nonignorable drop-out: selection models and mixture models. The two methods differ in how they factor the joint distribution of the complete response and the missing data vector. For selection models, the distribution is factored as

$$f(\mathbf{Y}_i, \mathbf{R}_i \mid \mathbf{X}_i, \gamma, \phi) = f_{\mathbf{Y}}(\mathbf{Y}_i \mid \mathbf{X}_i, \gamma) f_{\mathbf{R}|\mathbf{Y}}(\mathbf{R}_i \mid \mathbf{X}_i, \mathbf{Y}_i, \phi), \tag{3.1}$$

whereas for mixture models, the factorization is

$$f(\mathbf{Y}_i, \mathbf{R}_i \mid \mathbf{X}_i, \nu, \delta) = f_{\mathbf{R}}(\mathbf{R}_i \mid \mathbf{X}_i, \delta) f_{\mathbf{Y}|\mathbf{R}}(\mathbf{Y}_i \mid \mathbf{R}_i, \mathbf{X}_i, \nu). \tag{3.2}$$

Fitzmaurice and Laird (2000) discusses how selection models are attractive because they result in the desired distribution, $f_{\mathbf{Y}}$. Mixture models give $f_{\mathbf{Y}|\mathbf{R}}$, which then needs to be averaged over the distribution of drop-out patterns. Both types of models can suffer from identifiability concerns, but for selection models it is unclear how the restrictions that are necessary to alleviate this issue affect

the distribution of the missing responses. For mixture models, the affects from these restrictions are clear, and it is easier to determine when a model is identifiable.

Included under the selection and mixture models umbrella, Little (2009) makes further distinctions of these models including fixed-effects and mixed-effects models. The former are in the form of (3.1) and (3.2), and the latter group of models include random subject effects, $\mathbf{b}_i$. A mixed-effects selection model can be written as

$$f(\mathbf{Y}_i, \mathbf{R}_i, \mathbf{b}_i \mid \mathbf{X}_i, \gamma, \phi) = f_{\mathbf{B}}(\mathbf{b}_i \mid \mathbf{X}_i, \gamma_1) f_{\mathbf{Y}|\mathbf{B}}(\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{b}_i, \gamma_2) f_{\mathbf{R}|\mathbf{Y},\mathbf{B}}(\mathbf{R}_i \mid \mathbf{X}_i, \mathbf{Y}_i, \mathbf{b}_i, \phi) \quad (3.3)$$

and the mixed-effects mixture model is

$$f(\mathbf{Y}_i, \mathbf{R}_i, \mathbf{b}_i \mid \mathbf{X}_i, \nu, \delta) = f_{\mathbf{R}}(\mathbf{R}_i \mid \mathbf{X}_i, \delta) f_{\mathbf{B}|\mathbf{R}}(\mathbf{b}_i \mid \mathbf{X}_i, \mathbf{R_i}, \nu_1) f_{\mathbf{Y}|\mathbf{R},\mathbf{B}}(\mathbf{Y}_i \mid \mathbf{R}_i, \mathbf{X}_i, \mathbf{b}_i, \nu_2). \quad (3.4)$$

Little (2009) introduces mixed-effect hybrid models which combines attractive features of both selection and mixture models. This model is factored as

$$f(\mathbf{Y}_i, \mathbf{R}_i, \mathbf{b}_i \mid \mathbf{X}_i, \gamma, \delta, \nu) = f_{\mathbf{B}}(\mathbf{b}_i \mid \mathbf{X}_i, \gamma) f_{\mathbf{R}|\mathbf{B}}(\mathbf{R}_i \mid \mathbf{X}_i, \mathbf{b}_i, \delta) f_{\mathbf{Y}|\mathbf{R},\mathbf{B}}(\mathbf{Y}_i \mid \mathbf{R}_i, \mathbf{X}_i, \mathbf{b}_i, \nu). \quad (3.5)$$

When dealing with mixed-effects models, it is also important to differentiate between outcome-dependent drop-out and random-coefficient dependent drop-out models. Little (2009) writes the outcome-dependent drop-out mechanism as $f_{\mathbf{R}|\mathbf{Y},\mathbf{B}}(\mathbf{R}_i \mid \mathbf{X}_i, \mathbf{Y}_i, \mathbf{b}_i, \phi) = f_{\mathbf{R}|\mathbf{Y}}(\mathbf{R}_i \mid \mathbf{X}_i, \mathbf{Y}_i, \phi)$ and the random-coefficient dependent drop-out mechanism as $f_{\mathbf{R}|\mathbf{Y},\mathbf{B}}(\mathbf{R}_i \mid \mathbf{X}_i, \mathbf{Y}_i, \mathbf{b}_i, \phi) = f_{\mathbf{R}|\mathbf{B}}(\mathbf{R}_i \mid \mathbf{X}_i, \mathbf{b}_i, \phi)$ for selection models. The latter case gives the same formulation as shared-parameter models where both the response and missing data mechanism are dependent on the same latent variables. It is assumed that they are independent conditional on those latent variables. For pattern-mixture models, $f_{\mathbf{B}|\mathbf{R}}(\mathbf{b}_i \mid \mathbf{X}_i, \mathbf{R_i}, \nu_1) = f_B(\mathbf{b}_i \mid \mathbf{X}_i, \nu_1)$ for outcome-dependent drop-out and $f_{\mathbf{Y}|\mathbf{R},\mathbf{B}}(\mathbf{Y}_i \mid \mathbf{R}_i, \mathbf{X}_i, \mathbf{b}_i, \nu_2) = f_{\mathbf{Y}|\mathbf{B}}(\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{b}_i, \nu_2)$ for random-coefficient dependent drop-out.

Identifiability of model parameters is a major concern for parametric models, and some sort of restriction must be placed on the models to solve this issue. These restrictions are unverifiable from the observed data, and inference can be sensitive to these choices.

3.1 Selection Models

Selection models are of the form (3.1) or (3.3). Diggle and Kenward (1994) presents a fixed-effects model for multivariate normal responses, $\mathbf{Y}_i$. For a single subject, the authors first model a covariance structure for $\mathbf{Y}_i^* = (Y_{i1}^*, ..., Y_{in_i}^*)$, where $Y_{ij}^* = Y_{ij}$ if the subject has a response observed at time t_j and $Y_{ij}^* = 0$ if the response is missing.

The drop-out process is modeled using a logistic linear model, logit $\left\{P(D_i = t_j \mid D_i \geq t_j, \bar{Y}_{i,j})\right\} = \phi_0 + \phi_1 Y_{ij} + \sum_{k=2}^{j} \phi_k Y_{i(j+1-k)}$ where logit is defined as logit$(p) = \log(\frac{p}{1-p})$. This drop-out model depends on past observed data and the present possibly unobserved data, Y_{ij}, but not on any future unobserved responses. This is a common choice of model since it is believed that only past and current measurements would directly affect whether or not the patient drops out of the study. The model can be expanded by having the intercept, ϕ_0, depend on time-dependent covariates, or the model can be simplified by depending on fewer previous responses. The covariance and drop-out models combine to form the log-likelihood equation,

$$\begin{aligned} L(\gamma, \phi) &= L_1(\gamma_1, \gamma_2) + L_2(\phi) + L_3(\gamma, \phi) \\ L_1(\gamma_1, \gamma_2) &= \sum_{i=1}^{N} \log f_i^*(Y_{i1}, ...Y_{i(D_i-1)} \mid \gamma_1, \gamma_2) \\ L_2(\phi) &= \sum_{i=1}^{N} \sum_{k=2}^{D_i-1} \log\left\{1 - P(D_i = t_k \mid D_i \geq t_k, \bar{\mathbf{Y}}_{i(k-1)}, \phi)\right\} \\ L_3(\gamma, \phi) &= \sum_{i:D_i \leq n_i} \log P(D_i \mid \bar{\mathbf{Y}}_{i(D_i-1)}, \phi, \gamma), \end{aligned} \tag{3.6}$$

where f_i^* is the multivariate Gaussian distribution for the observed observations and $P(D_i \mid \bar{\mathbf{Y}}_{i,D_i-1}) = P(Y_{i,D_i}^* = 0 \mid Y_{i1}^* \neq 0, ..., Y_{i,D_i-1}^* \neq 0) = \int P(D_i = t_k \mid \bar{\mathbf{Y}}_{i(k-1)}, \phi) f_i^*(Y_k^* \mid Y_{i1}^*, ...Y_{i(k-1)}^*, \gamma_1, \gamma_2) dY_k^*$. L_1 is the log-likelihood for the distribution of all of the observed responses, L_2 is the log-likelihood representing the probabilities that a patient was observed at each time point up until they dropped out, and L_3 is the log-likelihood for the probability that the patient drops out when they do. Maximizing the likelihood is done using the simplex algorithm from Nelder and Mead (1965) in which no derivatives of the likelihood are necessary.

Fitzmaurice et al. (1996) works on a model for nonignorable drop-out when the responses are binary. They model the complete data likelihood as in Fitzmaurice and Laird (1993), $f(\mathbf{Y}_i \mid \gamma) = \exp\{\Psi_i'\mathbf{Y}_i + \Omega_i'\mathbf{W}_i - A(\Psi_i, \Omega_i)\}$, where $\gamma = (\Psi, \Omega)$, $\mathbf{W}_i = (Y_{i1}Y_{i2}, ..., Y_{i(n_i-1)}Y_{in_i}, ..., Y_{i1}Y_{i2}...Y_{in_i})$ is the vector of all interaction terms, and $A(\Psi_i, \Omega_i)$ is a normalizing constant. The observed likelihood is then $f(\mathbf{Y}_i^o, \mathbf{R}_i \mid \gamma, \phi) = \sum_{\mathbf{Y}_i^m} f(\mathbf{Y}_i \mid \gamma) f(\mathbf{R}_i \mid \mathbf{Y}_i, \phi)$. Similar to Diggle and Kenward (1994), Fitzmaurice et al. (1996) models the response (instead of drop-out) probability as a logistic regression model depending on the current and previous observations, $\text{logit}\left\{P(R_{ik} = 1 \mid R_{i1} = ... = R_{i(k-1)} = 1, \bar{\mathbf{Y}}_{ik})\right\} = \phi_0 + \phi_1 Y_{ik} + \sum_{j=2}^{k} \phi_j Y_{i(k+1-j)}$. Note that for this and the previous drop-out model, $\phi_1 = 0$ corresponds to MAR and $\phi_1 = ... = \phi_k = 0$ corresponds to MCAR. The model does not include covariates for simplicity, but they may be added in. Estimates of (γ, ϕ) are calculated using the EM algorithm as detailed in the paper.

Molenberghs et al. (1997) uses a similar model for the missingness mechanism, but applies a different response model using methods introduced by Molenberghs and Lesaffre (1994) in order to accommodate ordinal categorical responses.

Ibrahim et al. (2001) analyzes continuous responses with random effects, as in (3.3). Under the normal random effects model, $\mathbf{Y}_i = \mathbf{X}_i\gamma_2 + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i$, it is assumed that $\mathbf{b}_i \sim N(0, \mathbf{B}(\gamma_1))$ and $\epsilon_i \sim N(0, \sigma^2\mathbf{I})$ where $\mathbf{Z}_i$ is a design matrix of the covariates assumed to have random effects. Computational issues arise in the classical EM algorithm when the amount of missing observations or the number of random effects is large, so the Ibrahim paper introduces a variation of the Monte Carlo EM (MCEM) algorithm in which the random effects are integrated out of the log-likelihood. For the missing data mechanism, Ibrahim et al. (2001) models $P(R_{ij} = 1 \mid \mathbf{Y}_i, \phi)$ logistically which can include any terms from $\mathbf{Y}_i, \mathbf{X}_i,$ and $\mathbf{Z}_i$. Stubbendick and Ibrahim (2003) extends this analysis by allowing for nonignorable missing covariates as well as nonignorable missing continuous responses using a similar approach as Ibrahim et al. (2001). The model also allows for nonmonotone missingness, and can be solved using an MCEM method.

Stubbendick and Ibrahim (2006) models correlated discrete responses with generalized linear mixed models (GLMMs), $f(Y_{ij} \mid \mathbf{b}_i, \gamma_2, \tau) = \exp\left[\tau\left\{Y_{ij}\theta(\eta_{ij}) - g(\theta(\eta_{ij}))\right\} + c(Y_{ij}, \tau)\right]$, where τ is the dispersion parameter, $\theta(\cdot)$ is the link function, and $\eta_{ij} = \mathbf{X}'_{ij}\gamma_2 + \mathbf{Z}'_{ij}\mathbf{b}_i$. MCEM methods are used to solve the system and bootstrap methods can be used to obtain standard errors.

When using selection models, the first step is figuring out the proper multivariate model for the response. The models for the missing data mechanism tend to be logistic, with the choice of which responses and covariates to include in the model. Including too many variables can lead to identifiability issues. Once the full likelihood is specified, methods such as the EM algorithm can be used to solve for the desired parameters, but this algorithm does not provide standard errors which are necessary for confidence intervals or testing. Fitzmaurice et al. (1996) suggests using the empirical information from the last iteration of the EM algorithm or using the bootstrap method to obtain standard errors.

3.2 Mixture Models

Mixture models follow the form of (3.2) and (3.4) and model the distribution of the response conditional on the missingness pattern. This is why some refer to this formulation as pattern-mixture models (Little 1995). These models tend to be easy to fit and easy to determine identifiability. On the other hand, these models are often overspecified and the marginal, not conditional, distribution of the responses is usually desired which necessitates averaging over missingness patterns. Rewriting the distribution of responses for a patient who drops out prior to time $t_j = r$ as

$$\begin{aligned} f(\mathbf{Y}_i \mid D_i = r) &= f(\mathbf{Y}_i^o \mid D_i = r) f(\mathbf{Y}_i^m \mid \mathbf{Y}_i^o, D_i = r) \\ &= f(Y_{i1}, ..., Y_{i,r-1} \mid D_i = r) f(Y_{ir}, ..., Y_{in_i} \mid Y_{i1}, ..., Y_{i,r-1}, D_i = r), \end{aligned} \tag{3.7}$$

it is clear that the first distribution on the right hand side, the joint density function of the data through time point $r-1$, can be modeled from the observed data. The distribution $f(\mathbf{Y}_i^m \mid \mathbf{Y}_i^o, D_i = r)$ cannot be estimated from the data without restrictions. Little (1994) illustrates his proposed restrictions using an example of two bivariate normal observations, (Y_{i1}, Y_{i2}), with only

the second observation possibly missing. This model contains eleven parameters $\left\{\mu_1^{(r)}, \mu_2^{(r)}, \sigma_{11}^{(r)}, \sigma_{12}^{(r)}, \sigma_{22}^{(r)}, \pi; r = D_i - 1 = 1, 2\right\}$ with separate mean and covariance parameters for completers and those who drop out. The parameter $\pi = P(D_i = 2)$ is the probability of a completer. Three of the parameters, $\left\{\mu_2^{(2)}, \sigma_{12}^{(2)}, \sigma_{22}^{(2)}\right\}$, are unidentifiable from the observed data.

Little uses the drop-out model $P(D_i = 1 \mid \mathbf{Y}_i) = g(Y_{i1} + \lambda Y_{i2})$, where λ is known and $g(\cdot)$ is an unspecified function. The paper restrains that the conditional distribution of $Y_{i1} \mid Y_{i1} + \lambda Y_{i2}$ is the same across both missingness patterns (Little 1994, Little 1995). This is enough for all parameters to be identifiable without specifying $g(\cdot)$, unlike in selection models, and is called complete case missing values (CCMV) (Molenberghs 2004) .

It is evident that this restriction is highly dependent on choice of λ. When $\lambda = 0$, this corresponds to the MAR assumption since missingness is not dependent on any unobserved measurements. As λ increases in absolute value, the missingness depends more on possibly unobserved responses. It is unverifiable from the data whether or not $\lambda = 0$. Due to this dependence on the specification of λ, Little (1994) does a sensitivity analysis by choosing a few appropriate values and analyzing the data using each to see how the choice affects the analysis. Generally, values are chosen to be positive since that assumes that successive time points are positively correlated, but a negative λ might indicate that the difference in responses between time points would influence the drop-out. Little also suggests the possibility of choosing a prior distribution for λ and sampling from that when doing a Bayesian analysis.

Kenward (2003) gives an identifying restriction different from CCMV in Little (1994). The paper assumes that drop-out is only dependent on the past observed responses and current possibly unobserved response, not any future responses. This is called non-future missing values (NFMV) satisfying $f(Y_{it} \mid Y_{i1}, ..., Y_{i,t-1}, D_i = j) = f(Y_{it} \mid Y_{i1}, ..., Y_{i,t-1}, D_i \geq t - 1)$ for all $j < t - 1$ and $t \geq 2$. This means that the conditional density of the response of a subject at time t for those who have dropped out before time $t - 1$ can be modeled the same as the response of subjects who

have observations at least through time t. This restriction only leaves $f(Y_{i,t+1} \mid Y_{i1}, ..., Y_{it}, D_i = t)$ unidentified. The scientist can specify this distribution however he chooses. Kenward gives little direction in how to do this only saying, "substantive considerations can be used to identify this density, or a family of densities can be considered by way of sensitivity analysis."

Fitzmaurice and Laird (2000) propose the model $g\left(E\left[Y_{ij} \mid D_i, \mathbf{X}_{ij}\right]\right) = \mathbf{C}'_{ij}\nu$ where g is a known link function and $\mathbf{C}_{ij}$ can depend on both drop-out time and covariates. This model is appropriate for continuous and discrete data. Let π_l be the probability of a certain missingness pattern. This can be estimated by the sample proportion of subjects who drop out at each time point stratified by other covariates, like treatment, if desired in the model. For example, if there are two treatment groups and three possible times to drop out, then there are six different patterns and $l = 1, ..., 6$. For now, assume that the model is not stratified by covariates, but only by drop-out time. Then the marginal expectation can be solved for as $E(Y_{ij} \mid \mathbf{X}_{ij}) = \mu_{ij} = \sum_{l=1}^{n_i} \pi_l g^{-1}(\mathbf{C}'_{ij}\nu)$.

An issue with this solution is that the distributional form of the conditional mean, $E(Y_{ij} \mid \mathbf{X}_i, D_i)$, is assumed in formulating the model, but this distribution does not translate for the marginal mean, $E(Y_{ij} \mid \mathbf{X}_i)$, if the link is not linear (Little 2009).

Whereas Little (1994), makes assumptions to allow an over-specified model to be identified, Fitzmaurice and Laird (2000) only chooses models that are identifiable to begin with, which is easy to recognize in the mixture model framework. This often includes assuming similar patterns across drop-out times and the choice of model itself is an important assumption. For instance, when there are three time points, drop-out can occur at the second and third observations. Consider the model $g\left(E\left[Y_{ij} \mid D_i, \mathbf{X}_{ij}\right]\right) = \nu_0^{(r)} + \nu_1^{(r)} t_j$ with $r = 2, 3$ representing the time of drop-out and $r = 4$ indicating a completer. The parameter $\nu_1^{(2)}$ is not identifiable since a slope can't be estimated for those who drop-out at the second time point and therefore don't have at least two observations. On the other hand, the model $g\left(E\left[Y_{ij} \mid D_i, \mathbf{X}_{ij}\right]\right) = \nu_0^{(r)} + \nu_1 t_j$ $r = 2, 3, 4$ with a common slope across drop-out patterns is identifiable. The paper suggests performing a sensitivity analysis by

testing various identifiable models.

Fitzmaurice and Laird (2000) use generalized estimating equations (GEE) to produce estimates and variances for the parameters, ν which eliminates the need to specify the complete joint distribution of responses and the drop-out mechanism. The estimating equations are $\sum_{i=1}^{N} \mathbf{G}_i' \mathbf{V}_i^{-1} \left[\mathbf{Y}_i^o - E(\mathbf{Y}_i^o \mid D_i, \mathbf{X}_i, \nu)\right] = \mathbf{0}_{p\times 1}$, where ν is a $p \times 1$ vector of parameters, $\mathbf{G}_i = \partial E(\mathbf{Y}_i^o \mid D_i, \mathbf{X}_i, \nu)/\partial\nu$, and $\mathbf{V}_i$ is the working covariance matrix of the responses, $\mathbf{Y}_i^o$. Then $N^{1/2}(\hat{\nu} - \nu)$ is asymptotically normal with mean $\mathbf{0}$ and variance estimated by the sandwich variance estimator. The parameters ν from this model can easily be calculated using PROC GENMOD in SAS by including drop-out time and the interactions of drop-out time and other covariates in the model statement. The parameter estimates can estimate marginal means, $\hat{\mu}_{ij} = \sum_{l=1}^{n_i} \hat{\pi}_l g^{-1}(\mathbf{C}_{ij}'\hat{\nu})$, and standard errors can be obtained using bootstrap, jackknife, or the delta method.

3.3 Mixed-Effect Hybrid Models

Little (2009) introduced the concept of mixed-effect hybrid models (MEHMs) and Yuan and Little (2009) describes this model in more detail when $n_i = n$, or all subjects have the same number of planned visits. These models are desirable as they directly model the drop-out process like selection models while being computationally simple like mixture models. In some cases they can even be fit with PROC NLMIXED in SAS. The model in Yuan and Little (2009) differs slightly from the model in (3.5) in that it models drop-out time, D_i, instead of the vector of response indicators, $\mathbf{R}_i$: $f(\mathbf{Y}_i, D_i, \mathbf{b}_i \mid \mathbf{X}_i, \gamma, \delta) = f_{\mathbf{B}}(\mathbf{b}_i \mid \mathbf{X}_i, \gamma_1) f_{D|\mathbf{B}}(D_i \mid \mathbf{X}_i, \mathbf{b}_i, \delta) f_{\mathbf{Y}|D,\mathbf{B}}(\mathbf{Y}_i \mid D_i, \mathbf{X}_i, \mathbf{b}_i, \nu_2)$. Shared-parameter models are a subset of MEHMs where the response is modeled the same for different patterns of missingness, $f_{\mathbf{Y}|D,\mathbf{B}}(\mathbf{Y}_i \mid D_i, \mathbf{X}_i, \mathbf{b}_i, \nu_2) = f_{\mathbf{Y}|\mathbf{B}}(\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{b}_i, \nu_2)$.

The following linear mixed-effects model is used where $\mathbf{Z}_i$ is a known design matrix for the random effects, $\mathbf{b}_i$, while $\Sigma^{(j)}$ and Γ are unknown covariance matrices:

$$\begin{aligned} \mathbf{Y}_i \mid D_i = t_j, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i &\sim N_n(\mathbf{X}_i\beta^{(j)} + \mathbf{Z}_i\mathbf{b}_i, \Sigma^{(j)}) \\ \mathbf{b}_i \mid \Gamma &\sim N_q(\mathbf{0}, \Gamma). \end{aligned} \tag{3.8}$$

The discrete hazard rate is defined, $\lambda_{ij} = P(D_i = t_j \mid D_i \geq t_j; \mathbf{w}_{ij}, \mathbf{b}_i)$ where $\mathbf{w}_{ij}$ is some subset of fixed covariates. Note that $\lambda_{i(n+1)} = 1$. Then the missingness mechanism can be specified as $\pi_{ij} = P(D_i = t_j \mid \mathbf{w}_{ij}, \mathbf{b}_i) = \lambda_{ij}\prod_{k=1}^{j-1}(1-\lambda_{ik})$. Averaging over missingness pattern and random effects, the marginal mean response is $E(\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{Z}_i) = E_{\mathbf{b}_i}E_{D_i|\mathbf{b}_i}\left(\mathbf{X}_i\beta^{(j)} + \mathbf{Z}_i\mathbf{b}_i\right) = \mathbf{X}_i\sum_{j=1}^{n} E_{\mathbf{b}_i}(\pi_{ij})\beta^{(j)}$ where $E_{\mathbf{b}_i}(\pi_{ij})$ can be estimated by the sample proportion of subjects with each missingness pattern if the drop-out mechanism is assumed independent of the random effects. Otherwise, the expectation can be calculated numerically by integrating the random effects out of the drop-out model.

This model is not identifiable without additional assumptions. If it is assumed the responses are independent conditioned on the random effects, then the restriction $\Sigma^{(j)} = \sigma_j^2 I$ can be placed on the model. If the responses are identically independently distributed given the random effects, then the restriction is $\Sigma^{(j)} = \sigma^2 I$. It can also be assumed that the covariance matrix is the same across missingness patterns, leading to $\Sigma^{(j)} = \Sigma$. If the investigators are willing to place these restrictions, then the variance parameters may be identifiable.

When estimating $\beta^{(j)}$, if the covariates are not time dependent, then the number of subjects in each missingness pattern needs to be greater than the number of parameters in order to be identifiable. If the covariates are time-dependent, then slopes are not identifiable. The cases that cause this issue can be eliminated, as these subjects have only had one visit and then dropped out.

To estimate the parameters in (3.8), Yuan and Little (2009) maximize the marginal likelihood, integrated over the random effects and missing data: $L(\theta \mid \mathbf{D}, \mathbf{Y}^o, \mathbf{X}) = \prod_{i=1}^{N}\int\int f(\mathbf{Y}_i, D_i, \mathbf{b}_i \mid \mathbf{X}_i, \theta)d\mathbf{Y}_i^m d\mathbf{b}_i$. The integration is done using adaptive Gaussian quadrature approximation and the maximization is done using a quasi-Newton approach, both available in PROC NLMIXED.

4 Identifiability

Fitzmaurice et al. (1996) states that if there exist parameters $(\gamma_1, \phi_1) \neq (\gamma_2, \phi_2)$ such that $f(\mathbf{Y}_i^o, \mathbf{R}_i \mid \gamma_1, \phi_1) = f(\mathbf{Y}_i^o, \mathbf{R}_i \mid \gamma_2, \phi_2)$ then the parameters are not identifiable. Some identifiability results have been presented for mixture and mixed-effect hybrid models, but little detail has been given for selection models, as it is difficult to determine the identifiability (Fitzmaurice and Laird 2000).

Tang et al. (2003) gives theorems and properties relating to model identifiability. The paper proposes borrowing methods from the area of response-biased sampling which gets its estimates from maximizing the conditional likelihood of $\mathbf{X} \mid \mathbf{Y}$ instead of $\mathbf{Y} \mid \mathbf{X}$. The paper assumes that $\mathbf{X}$ is fully observed and only responses are missing. The joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ can be factored as $f(\mathbf{X}_i, \mathbf{Y}_i \mid \xi) = f_{\mathbf{X}}(\mathbf{X}_i \mid \xi_1) f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}_i \mid \mathbf{X}_i, \xi_2)$, and the drop-out model, $P(\mathbf{R}_i = \mathbf{r} \mid \mathbf{Y}_i, \mathbf{X}_i, \phi) = w(\mathbf{Y}_i, \phi, \mathbf{r})$, is only dependent on the responses, not covariates. This model is appropriate when it is believed that the responses can fully capture the patterns for drop-out, or in other words, drop-out and the covariates are independent given the responses. The likelihood is

$$\begin{aligned} L(\xi_1, \xi_2, \phi) &= \prod_{i=1}^{N} f_{\mathbf{X}}(\mathbf{X}_i \mid \xi_1) \int f(\mathbf{Y}_i, \mathbf{R}_i \mid \mathbf{X}_i, \xi_1, \xi_2, \phi) d\mathbf{Y}_i^m \\ &= \left[\prod_{i=1}^{N} f_{\mathbf{X}}(\mathbf{X}_i \mid \xi_1) \right] \left[\prod_{\mathbf{R}_i = \mathbf{1}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}_i \mid \mathbf{X}_i, \xi_2) w(\mathbf{Y}_i, \phi, \mathbf{r} = \mathbf{1}) \right] \\ &\times \left[\prod_{\mathbf{R}_i \neq \mathbf{1}} \int f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}_i \mid \mathbf{X}_i, \xi_2) w(\mathbf{Y}_i, \phi, \mathbf{r}) d\mathbf{Y}_i^m \right]. \end{aligned} \tag{4.1}$$

Let $L_1(\xi_1) \equiv \prod_{i=1}^{N} f_{\mathbf{X}}(\mathbf{X}_i \mid \xi_1)$. Then $\hat{\xi}_1$ can be obtained by maximizing $L_1(\xi_1)$ or using the empirical distribution, $F_N(\mathbf{X})$. Plugging in this estimate, $\hat{\xi}_2$ can be estimated by maximizing

$$L_2(\xi_2, F_N(\mathbf{X})) = \prod_{\mathbf{R}_i = \mathbf{1}} \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}_i \mid \mathbf{X}_i, \xi_2)}{\int f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}_i \mid \mathbf{X}_i, \xi_2) dF_N}. \tag{4.2}$$

This gives the parameters of interest. All observations are used the obtain $\hat{\xi}_1$, but only the completers are used to directly calculate $\hat{\xi}_2$. This estimate is still dependent on the drop-outs through $\hat{\xi}_1$. This is not as efficient as maximum likelihood methods when the drop-out mechanism is correctly

specified, but this method doesn't require the mechanism to be specified. Hence this method is more robust when the mechanism is unknown. The theoretical properties about the identifiability of these parameters are given in Tang et al. (2003).

Shao and Zhao (2013) uses similar pseudolikelihood estimators, but allows for more flexible assumptions. The paper specifically states that for their model to be identifiable at least one of $(Y_{i1}, ..., Y_{iD_i}, \mathbf{X}_i)$ must not be related to drop-out conditional on the other components. The covariates can be partitioned $\mathbf{X} = (\mathbf{U}, \mathbf{Z})$ where $\mathbf{U}$ is related to drop-out while $\mathbf{Z}$ is not related to drop-out conditional on $\mathbf{U}$ and the responses. The assumption about the important covariates that make up $\mathbf{U}$ cannot be tested, so again the investigator needs to do a sensitivity analysis.

5 Semiparametric Methods

Rotnitzky et al. (1998) and Scharfstein et al. (1999) introduce semiparametric methods to analyze nonignorable drop-outs. These models encompass a class of possible models attempting to have less untestable restrictive assumptions.

5.1 Parametric Missing Data Mechanism

Rotnitzky et al. (1998) introduces augmented inverse probability of censoring weighted (AIPCW) estimators of a marginal mean of responses at a fixed time. This estimator is consistent and asymptotically normal as long as the drop-out model is correctly specified. The estimator can be solved as the solution to the AIPCW estimating equations

$$\begin{aligned} &\sum_{i=1}^{N} \left[\frac{I(\mathbf{R}_i = \mathbf{1})}{\pi_i(\mathbf{1})} d(\mathbf{X}_i, \beta) \left\{ \mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{X}_i, \beta) \right\} + A_i \right] = 0, \\ &\text{where } A_i = \sum_{\mathbf{r} \neq \mathbf{1}} \left[I(\mathbf{R}_i = \mathbf{r}) - \frac{I(\mathbf{R}_i = \mathbf{1})}{\pi_i(\mathbf{1})} \pi_i(\mathbf{r}) \right] \phi_{\mathbf{r}}(\bar{\mathbf{Y}}_{(\mathbf{r})i}). \end{aligned} \tag{5.1}$$

The function $E(\mathbf{Y}_i \mid \mathbf{X}_i, \beta)$ is specified, $\pi_i(\mathbf{r}) = P(\mathbf{R}_i = \mathbf{r} \mid \mathbf{Y}_i)$, and $d(\mathbf{X}_i, \beta)$ is a matrix of fixed functions, a common choice being $d(\mathbf{X}_i, \beta) = \frac{\partial E(\mathbf{Y}_i | \mathbf{X}_i, \beta)}{\partial \beta} \text{Var}(\mathbf{Y}_i \mid \mathbf{X}_i)$. Also, $\phi_{\mathbf{r}}(\mathbf{Y}_{(\mathbf{r})i})$ is an arbitrary vector of functions of the data observed for missingness pattern $\mathbf{r}$.

The first part of the estimating equation only uses data from the completers and the augmentation part, A_i, uses the observed data from all subjects. The solution $\hat{\beta}$ is asymptotically normal and the variance can be estimated with the sandwich variance estimate. Since $\hat{\beta}$ depends on the choice of $\phi_{\mathbf{r}}(\mathbf{Y}_{(\mathbf{r})i})$, efficiency can be improved according to Rotnitzky et al. (1998) by choosing $\tilde{\phi}_{\mathbf{r}}(\mathbf{Y}_{(\mathbf{r})i}) = -C\phi_{\mathbf{r}}(\mathbf{Y}_{(\mathbf{r})i})$ where $C = E\left[\left\{\frac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{1})} d(\mathbf{X}_i,\beta)\left\{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{X}_i,\beta)\right\}\right\} A_i'\right] \text{Var}(A_i)^{-1}$ which can be estimated by sample averages.

When the probability of each missingness pattern, $\pi_i(\mathbf{r})$, needs to be estimated, it can be modeled logistically or as chosen, and the following estimating equation can be solved:

$$\sum_{i=1}^{N} A_i(\alpha) = \sum_{i=1}^{N} \sum_{\mathbf{r}\neq\mathbf{1}} \left[I(\mathbf{R}_i = \mathbf{r}) - \frac{I(\mathbf{R}_i = \mathbf{1})}{\pi_i(\mathbf{1},\alpha)} \pi_i(\mathbf{r},\alpha) \right] \phi_{\mathbf{r}}(\bar{\mathbf{Y}}_{(\mathbf{r})i}) = 0. \tag{5.2}$$

This is necessary to estimate when the missing data mechanism is not by design. The two estimating equations can be stacked and then solved together to estimate $\hat{\alpha}$ and $\hat{\beta}$ simultaneously.

5.2 Semiparametric Missing Data Mechanism

Scharfstein et al. (2009) broadens this model so the drop-out mechanism is also semiparametric, specified as a Cox proportional hazards model with the form $\lambda_D(t \mid \bar{\mathbf{X}}_t, \mathbf{Y}) = \lim_{h\to 0} P(t \leq D < t+h \mid \bar{\mathbf{X}}_t, \mathbf{Y}, D \geq t)/h = \lambda_0(t \mid \bar{\mathbf{X}}_t)\text{exp}(\alpha_0 \mathbf{Y})$. This indicates that drop-out can be modeled continuously and is only dependent on the unobserved responses in the term $\text{exp}(\alpha_0\mathbf{Y})$. For fixed α_0 and $F_{\mathbf{Y}^o}$ of the observed data, Scharfstein et al. (1999) claims that there is a unique $\lambda_0(t \mid \bar{\mathbf{X}}_t)$ and unique joint distribution $F_{\bar{\mathbf{X}}_t,\mathbf{Y},D}$ for this data. This extends to the case where $\alpha_0\mathbf{Y}$ is replaced by a function, r, which can depend on time, α_0, covariates, and responses. Because this result is only valid when α_0 is known, the authors suggest doing a sensitivity analysis for a range of different values. When a decision needs to be made, the authors suggest a nonparametric Bayesian analysis by putting a prior distribution on α_0 and the form of $r(t, \alpha_0, \bar{\mathbf{X}}(t_{n_i}), \mathbf{Y})$.

Similar to when the missing mechanism is parametrically modeled, Scharfstein et al. (1999) solves for the conditional mean response $\hat{\mu}$ using estimating equations,

$$\sum_{i=1}^{N} \frac{\Delta}{\hat{\pi}(\mathbf{R}_i = \mathbf{1}, \bar{\mathbf{X}}_{in})} \left\{ \mathbf{Y}_i - \mu - \hat{E}\left[(1-\Delta) b(\bar{\mathbf{X}}_{iD_i}, D_i, \mu) \mid \bar{\mathbf{X}}_{in}, \mathbf{Y}_i \right] \right\} + (1-\Delta) b(\bar{\mathbf{X}}_{iD_i}, D_i \mid \mu) = 0 \tag{5.3}$$

where $\Delta = I(\mathbf{R}_i = \mathbf{1})$, b is a function chosen by the investigator, and $\pi(\mathbf{R}_i = \mathbf{1}, \bar{\mathbf{X}}_{in}) = P(\mathbf{R}_i = \mathbf{1} \mid \mathbf{X}_i, \mathbf{Y}_i)$ which can now be dependent on the covariates. The function $\Lambda(t \mid \bar{\mathbf{X}}_t) = \int_0^t \lambda_0(u \mid \bar{\mathbf{X}}_u)$ is the cumulative conditional baseline hazard function. It can be shown that $\hat{\pi}(\mathbf{R}_i = \mathbf{1}, \bar{\mathbf{X}}_{in}) = \exp\left[-\hat{\Lambda}(t_n \mid \bar{\mathbf{X}}_{in}) \exp(\alpha_0 \mathbf{Y}) \right]$ and

$$\hat{E}\left[(1-\Delta) b(\bar{\mathbf{X}}_{iD_i}, D_i, \mu) \mid \bar{\mathbf{X}}_{in}, \mathbf{Y}_i \right] = \int_0^n b(\bar{\mathbf{X}}_{it}, t, \mu) \exp\left[-\hat{\Lambda}(t_n \mid \bar{\mathbf{X}}_{in}) \exp(\alpha_0 \mathbf{Y}) \right] \times \exp(\alpha_0 \mathbf{Y}) d\hat{\Lambda}(t \mid \bar{\mathbf{X}}_{it}). \tag{5.4}$$

It is assumed that α_0 is fixed, so this leaves only the estimation of $\Lambda(t \mid \bar{\mathbf{X}}_t) = \int_0^t \lambda_0(u \mid \bar{\mathbf{X}}_u)$ as the only unknown in (5.3). When there is no missing data, the Nelson-Aalen estimator is popular for estimating the cumulative baseline hazard function. There is an identifiability issue if the covariates have jumps at many time points or if there are multiple components which are continuous. To help alleviate this issue, it is assumed that the covariates are constant over time, $\bar{\mathbf{X}}_t = \mathbf{X}$. If $\mathbf{X}$ is univariate and continuous, values can be binned to make it discrete. When $\mathbf{X}$ is discrete, then $\Lambda(t \mid \mathbf{X})$ can be estimated with the Nelson-Aalen estimator at each level of $\mathbf{X}$.

Letting $n_{\mathbf{x}} = \sum_{i=1}^{N} I(\mathbf{X}_i = \mathbf{x})$ be the number of subjects with covariates $\mathbf{X} = \mathbf{x}$ and $dN_i^{\mathbf{x}}(u) = I(\mathbf{X}_i = \mathbf{x}, D_i \leq u, \Delta_i = 0)$ be the indicator that a non-completer with covariates $\mathbf{X} = \mathbf{x}$ dropped out before time u, the Nelson-Aalen estimator that accounts for the missing data is

$$\hat{\Lambda}(t \mid \mathbf{X} = \mathbf{x}) = \int_0^t \left(\frac{1}{n_{\mathbf{x}}} \sum_{i=1}^{n} \frac{\Delta_i I(\mathbf{X}_i = \mathbf{x}) \exp(\alpha_0 \mathbf{Y}_i) I(D_i \geq u)}{\exp\left\{ -\exp(\alpha_0 \mathbf{Y}_i) \left[\hat{\Lambda}(t_n \mid \mathbf{X}_i = \mathbf{x}) - \hat{\Lambda}(u \mid \mathbf{X}_i = \mathbf{x}) \right] \right\}} \right)^{-1} \times \left(\frac{1}{n_{\mathbf{x}}} \sum_{i=1}^{N} dN_i^{\mathbf{x}}(u) \right). \tag{5.5}$$

This estimator is a step function jumping at each unique drop-out time, and needs to be obtained

recursively. Denote the jump sizes as $Q^{\mathbf{x}}_{(1)} < ... < Q^{\mathbf{x}}_{(k_{\mathbf{x}})}$ where $k_{\mathbf{x}}$ is the number of jumps for $\mathbf{X} = \mathbf{x}$, and denote the number of subjects who drop-out at $Q^{\mathbf{x}}_{(k)}$ as $c^{\mathbf{x}}_k$. The following steps are necessary to obtain the estimator $\hat{\Lambda}(t \mid \mathbf{X} = \mathbf{x})$:

1. Compute $\hat{\lambda}^{\mathbf{x}}_{(k_{\mathbf{x}})} = \left[\sum_{i=1}^{N} \Delta_i I(\mathbf{X}_i = \mathbf{x}) \exp(\alpha_0 \mathbf{Y}_i)\right]^{-1} c^{\mathbf{x}}_{k_{\mathbf{x}}}$.

2. Recursively compute for $k = k_{\mathbf{x}} - 1, ..., 1$: $\hat{\lambda}^{\mathbf{x}}_k = \left(\sum_{i=1}^{N} \frac{\Delta_i I(\mathbf{X}_i = \mathbf{x}) \exp(\alpha_0 \mathbf{Y}_i)}{\exp\left\{-\exp(\alpha_0 \mathbf{Y}_i) \sum_{j=k+1}^{k_{\mathbf{x}}} \hat{\lambda^{\mathbf{x}}_j}\right\}}\right)^{-1} c^{\mathbf{x}}_k$.

3. Calculate the estimate $\hat{\Lambda}(t \mid \mathbf{X} = \mathbf{x}) = \sum_{k=1}^{k_{\mathbf{x}}} \hat{\lambda}^{\mathbf{x}}_k I(Q_{(k)} \leq t)$.

Now $\hat{\mu}(b)$ can be obtained from the estimating equations (5.3). This estimator is regular and asymptotically linear (RAL) with asymptotic variance equal to the semiparametric variance bound for the model. This variance can be estimated by $N^{-1}\tilde{\tau}(b^*)^{-2} \sum_{i=1}^{N} h(\text{observed data}, \hat{\mu}(b), \hat{\Delta}, b^*)$ where h is the summand in (5.3) and $\tilde{\tau}(b)$ is a consistent estimator of $\tau(b)$, the first derivative of the expected value of the estimating equation evaluated at μ_0.

Scharfstein et al. (1999) shows that using the property that for any function l, $\hat{E}(l(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}) = \frac{1}{n_{\mathbf{x}}} \sum_{i=1}^{N} \frac{\Delta_i I(\mathbf{X}_i = \mathbf{x}) l(\mathbf{Y}_i)}{\hat{\pi}(\mathbf{x}, \mathbf{Y}_i)}$, b^* can be estimated by $\hat{b^*} = \frac{\hat{E}[(\mathbf{Y} - \mu) \exp(\alpha_0 \mathbf{Y}) | \mathbf{X} = \mathbf{x}]}{\hat{E}[\exp(\alpha_0 \mathbf{Y}) | \mathbf{X} = \mathbf{x}]}$. The estimator $\hat{\mu}(\hat{b^*})$ is also RAL and takes the form

$$\hat{\mu}(\hat{b^*}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\Delta_i}{\hat{\pi}(\mathbf{X}_i, \mathbf{Y}_i)} \mathbf{Y}_i - \frac{\Delta_i - \hat{\pi}(\mathbf{X}_i, \mathbf{Y}_i)}{\hat{\pi}(\mathbf{X}_i, \mathbf{Y}_i)} \frac{\hat{E}\left[\mathbf{Y} \exp(\alpha_0 \mathbf{Y}) \mid \mathbf{X} = \mathbf{X}_i\right]}{\hat{E}\left[\exp(\alpha_0 \mathbf{Y}) \mid \mathbf{X} = \mathbf{X}_i\right]}. \tag{5.6}$$

It is an issue to find consistent asymptotically normal (CAN) estimators when $\bar{\mathbf{X}}_t$ is high dimensional, defined as having "two or more continuous components or many discrete components...,or $\bar{\mathbf{X}}_t$ jumps at many different times" (Scharfstein et al. 1999). Due to this curse of high dimensionality, Scharfstein et al. (1999) considers another similar model, but using $W_t = w(t, \bar{\mathbf{X}}_t)$ where w is a known function that reduces the dimension. This new model can be written as $\lambda_D(t \mid \bar{\mathbf{X}}_t, \mathbf{Y}) = \lambda_0(t)\exp(\gamma_0' W_t)$. Once again, the authors suggest a sensitivity analysis for the forms of both $w(t, \bar{\mathbf{X}}_t)$ and $r(t, \gamma_0, \bar{\mathbf{X}}_n, \mathbf{Y})$. The details for solutions for this model are similar to the previous model and can be found in Scharfstein et al. (1999).

6 Sensitivity Analysis

As stated previously, the assumption that data is missing nonignorably versus MAR cannot be verified by the observed data. The same is true for many model assumptions used to alleviate identifiability issues. For this reason, many authors encourage a sensitivity analysis. Kenward et al. (2001) distinguishes the difference between imprecision and ignorance which combine to form the total uncertainty. Imprecision is analyzed with standard errors and confidence intervals, while sensitivity analyses study ignorance that comes from having missing data. Kenward's paper discusses how to form ignorance intervals by having a class of models that would be identifiable with complete data and choosing a sensitivity parameter to fix. Then the investigator can vary the sensitivity parameter to see how that affects the estimable quantities of interest.

This idea is similar to previous methods discussed in this review paper. For the semiparametric methods in the previous section, a class of estimators was identified, and then the drop-out process was parameterized with fixed α_0. A sensitivity analysis can be performed by evaluating the results for different α_0 in a range of plausible values, similar to what Little (1995) suggests for λ in his mixture model restriction. Scharfstein et al. (2013) refers to this as a global sensitivity analysis.

An ad hoc approach would be to try different methods with different assumptions on the data and see how the outcome changes, as Fitzmaurice and Laird (2000) suggest for their mixture model. Another method for sensitivity analyses is a local approach where small deviations of a parameter, usually around MAR, are used to see the influence of that change on the analysis.

6.1 Normal Curvature

Verbeke et al. (2001) proposes a method based on local influence for selection models with multivariate Gaussian response as in Diggle (1994). The drop-out model, logit $\left\{P(D_i = j \mid \bar{\mathbf{Y}}_{i,j-1})\right\} = \phi_0 + \phi_1 Y_{ij} + \sum_{k=2}^{j} \phi_k Y_{i(j+1-k)}$, is perturbed slightly by including a subject-specific coefficient of the current response, logit $\left\{P(D_i = j \mid \bar{\mathbf{Y}}_{i,j-1})\right\} = \phi_0 + \phi_1^i Y_{ij} + \sum_{k=2}^{j} \phi_k Y_{i(j+1-k)}$. For MAR data,

$\phi_1 = 0$, so ϕ_1^i are seen as deviations from this mechanism. If a subject's ϕ_1^i causes a large difference in parameter estimates, then that subject has a big influence on the results. The authors try to capture this effect of nonignorable nonresponse by using local influence.

Denote $\theta = (\gamma, \phi_0, \phi_2, ...\phi_{n_i})$ as the parameters from the data and drop-out models, not including the coefficient of the current response, ϕ_1. Then the log-likelihood is $\ell(\theta \mid \boldsymbol{\phi}_1) = \sum_{i=1}^{N} \ell_i(\theta \mid \phi_1^i)$ where $\boldsymbol{\phi}_1 = (\phi_1^1, ..., \phi_1^N)$. Let $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_1^0 \equiv (0, ..., 0)$ represent the MAR case, and the maximum likelihood estimators are denoted $\hat{\theta}$ under MAR and $\hat{\theta}_{\boldsymbol{\phi}_1}$ with no restrictions.

Cook (1986) first proposed examining the graph of likelihood displacement, $\text{LD}(\boldsymbol{\phi}_1) = 2\left[\ell(\hat{\theta} \mid \boldsymbol{\phi}_1^0) - \ell(\hat{\theta}_{\boldsymbol{\phi}_1} \mid \boldsymbol{\phi}_1)\right]$ against $\boldsymbol{\phi}_1$. Since this can only be visualized in two dimensions, Cook (1986) looks at the normal curvature, or $C_{\mathbf{h}} = 2|\mathbf{h}'\Psi'\ddot{L}^{-1}\Psi\mathbf{h}|$, where $\ddot{L} = \frac{\partial^2 \ell(\theta|\phi^0)}{\partial\theta'\partial\theta}$ and $\Psi_i = \frac{\partial^2 \ell_i(\theta|\phi_1^i)}{\partial\phi_1^i\partial\theta}\big|_{\theta=\hat{\theta},\phi_1^i=0}$. The local influence measure for each subject can be calculated if $\mathbf{h}$ is the $\mathbf{0}$-vector with 1 in the ith position. This is simplified as $C_i = 2|\Psi_i'\ddot{L}^{-1}\Psi_i|$. When this value is high, that means that this subject has a large effect on inference if they had dropped out nonrandomly while all others had dropped out randomly.

If certain subjects have a high local influence, they can be deleted or have their responses increased in magnitude for a subsequent analysis to research further how that affects the outcome. Verbeke et al. (2001) notes that this measurement of influence does not show how strong the nonignorability is in the data, but rather is just a tool for sensitivity analysis.

6.2 Index of Local Sensitivity to Nonignorability (ISNI)

Xie (2008) has a different approach to local influence. His paper extends the index of local sensitivity to nonignorability (ISNI) introduced by Ma et al. (2005) to the non-Gaussian case. The missing data mechanism is assumed to be similar to that in Verbeke et al. (2001), but allows for functions other than logit and the dependence on current covariates. The parameter ϕ_1 remains the coefficient of the current unobserved observation Y_{ij}. The MLEs for the data likelihood are computed for the two situations when $\phi_1 = 0$ and for when ϕ_1 is fixed. These MLEs are denoted

$\hat{\gamma}(0)$ and $\hat{\gamma}(\phi_1)$ respectively. Then the approximation, $\hat{\gamma}(\phi_1) \approx \hat{\gamma}(0) + \phi_1 \frac{\partial \hat{\gamma}(\phi_1)}{\partial \phi_1}\big|_{\phi_1=0}$ is given and $\frac{\partial \hat{\gamma}(\phi_1)}{\partial \phi_1}\big|_{\phi_1=0}$ is defined to be the ISNI. Xie (2008) gives the long formulas to calculate this quantity. This performs well as an approximation to $\hat{\gamma}(\phi_1)$, making it computationally easier to evaluate the model for different quantities of ϕ_1.

While ISNI doesn't measure the strength of nonignorability, the c statistic, $c = |\text{SE}/\text{ISNI}|$, aims to capture this information. The SE is the standard error of the MAR estimate of ϕ_1. The c statistic "denotes the critical value of ϕ_1 above which the bias due to nonignorable drop-out is larger than the sampling error" (Xie 2008). A small c value means that the parameter is locally sensitive to nonignorable drop-out and a large value means that the parameter is more robust to moderate drop-out.

Xie (2008) extends the definition of ISNI to the case where the drop-out mechanism differs for groups. In the case where there are two groups, there are separate parameters $\phi_1 = (\phi_{11}, \phi_{12})$ for coefficients of the current response and the dimension of ϕ_1 is $q = 2$. The extended ISNI aims to get at the maximum possible change of parameters by looking within a ball of radius $\sqrt{q}$ around the MAR case, $\phi_1 = 0$: $\text{ISNI}(\hat{\gamma}) \approx \max_{||\phi_1||=\sqrt{q}} [\hat{\gamma}(\phi_1) - \hat{\gamma}(0)] = \sqrt{q} \left\| \frac{\partial \hat{\gamma}(\phi_1)}{\partial \phi_1} \right\|^{1/2}_{\hat{\gamma}(0),\hat{\phi}_0(0),\phi_1=0}$. Both ISNI and normal curvature are useful tools in sensitivity analysis, but don't need to be used exclusively. The investigator can choose from many ways to evaluate the sensitivity of his model assumptions.

7 Conclusion

Selection, mixture, and semiparametric models have been proposed for dealing with nonignorable drop-out. A common theme is that parameters are not identifiable unless untestable assumptions are made. A sensitivity analysis is key in evaluating how the assumption of missing data mechanism affects estimates. The sensitivity analysis methods still seem ad hoc, so care needs to be taken when evaluating models. There doesn't appear to be any papers comparing how these methods perform under different assumptions, so this would be good for future work.

References

Cook, R. D., 1986: Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133–169.

Diggle, P. and M. G. Kenward, 1994: Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **43 (1)**, 49–93.

Fitzmaurice, G. M. and N. M. Laird, 1993: A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–151.

Fitzmaurice, G. M. and N. M. Laird, 2000: Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics*, **1**, 141–156.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware, 2011: *Applied Longitudinal Analysis*. 2d ed., John Wiley and Sons Inc.

Fitzmaurice, G. M., N. M. Laird, and G. E. P. Zahner, 1996: Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, **91 (433)**, 99–108.

Ibrahim, J. G., M. H. Chen, and S. R. Lipsitz, 2001: Missing response in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, **88**, 551–564.

Kenward, M. G., E. J. Goetghebeur, and G. Molenberghs, 2001: Sensitivity analysis for incomplete categorical data. *Statistical Modelling*, **1 (1)**, 31–48.

Kenward, M. G., G. Molenberghs, and H. Thijs, 2003: Pattern-mixture models with proper time dependence. *Biometrika*, **90 (1)**.

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

Liang, K. Y. and S. L. Zeger, 1986: Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 1322.

Little, R. J. A., 1994: A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81 (3)**, 471–483.

Little, R. J. A., 1995: Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Association*, **90**, 1112–1121.

Little, R. J. A. and D. Rubin, 2002: *Statistical Analysis with Missing Data*. 2d ed., John Wiley and Sons Inc.

Ma, G., A. B. Troxel, and D. F. Heitjan, 2005: An index of local sensitivity to nonignorable drop-out in longitudinal modelling. *Statistics in Medicine*, **24 (14)**, 2129–2150.

Molenberghs, G. and G. Fitzmaurice, 2009: Incomplete data: Inroduction and overview. *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, Eds., Chapman and Hall/CRC Press.

Molenberghs, G. and M. G. Kenward, 2007: *Missing data in clinical studies*. John Wiley and Sons Inc.

Molenberghs, G., M. G. Kenward, and E. Lesaffre, 1997: The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, **84**, 33–44.

Molenberghs, G. and E. Lesaffre, 1994: Marginal modelling of correlated ordinal data using an n-way plackett distribution. *Journal of the American Statitical Association*, **89**, 633–644.

Nelder, J. A. and R. Mead, 1965: A simplex method for function minimization. *The Computer Journal*, **7**, 303–313.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein, 1998: Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, **93 (444)**, 1321–1339.

Rubin, D. B., 1976: Inference and missing data. *Biometrika*, **63 (3)**, 581–592.

Scharfstein, D. O., A. McDermott, W. Olson, and F. Wiegand, 2013: Global sensitivity analysis for repeated measures studies with informative drop-out: A fully parametric approach. *Statistics in Biopharmaceutical Research*, submitted.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins, 1999: Adjusting for non-ignorable drop-out using semiparametric nonresponse models. *Journal of the American Association*, **94**, 1096–1120.

Shao, J. and J. Zhao, 2013: Estimation in longitudinal studies with nonignorable dropout. *Statistics and Its Interface*, **6**, 303–313.

Stubbendick, A. L. and J. G. Ibrahim, 2003: Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, **59 (4)**, 1140–1150.

Stubbendick, A. L. and J. G. Ibrahim, 2006: Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, **16**, 1143–1167.

Tang, G., R. J. A. Little, and T. E. Raghunathan, 2003: Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, **90 (4)**, 747–764.

Verbeke, G. and et al., 2001: Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics*, **57**, 7–14.

Xie, H., 2008: A local sensitivity analysis approach to longitudinal non-gaussian data with non-ignorable dropout. *Statistics in Medicine*, **27**, 3155–3177.

Yuan, Y. and R. J. A. Little, 2009: Mixed-effect hybrid models for longitudinal data with nonignorable dropout. *Biometrics*, **65 (2)**, 478–486.