Methods for Handling Dropouts in Longitudinal Studies

Garrett M. Fitzmaurice

Division of General Medicine, Brigham and Women's Hospital
Department of Biostatistics, Harvard School of Public Health

Joint work with Ken Wilkins

Outline

- Statement of Problem
- Motivating Example: Clinical trial of contracepting women
- Notation
- Selection and Pattern Mixture Models
- Marginally-Specified Pattern Mixture Models
- Concluding Remarks

Dropout in Longitudinal Studies

Most longitudinal studies are designed to collect data on every individual at each time of follow-up.

Commonly, not all responses are observed at all occasions.

Results in a large class of distinct missingness patterns.

Longitudinal studies frequently suffer from dropout:

Some individuals “drop-out” of study before intended completion time and thus have incomplete responses.

Reasons for dropout: happenstance, adverse events, lack of efficacy.

Methods currently available via commercial software assume (at best) that dropout is “ignorable”.

When dropout is “ignorable”, probability of dropout does not depend upon the unobserved events (Rubin, 1976).

When probability of dropout depends upon the unobserved events it is said to be “nonignorable”.

If dropout is “nonignorable”, bias can potentially arise.

Need for simple methods that can handle “nonignorable” dropout.

Example: Clinical trial of contracepting women

Randomized clinical trial comparing two doses of a contraceptive:
100 mg or 150 mg of DMPA, given at 90-day intervals.

Woman completed a menstrual diary that recorded any vaginal bleeding pattern disturbances.

Outcome of interest is a repeated binary response indicating whether or not a woman experienced amenorrhea (absence of menstrual bleeding).

A total of 1151 women completed the menstrual diaries.

Dropout: There was substantial dropout for reasons that were thought likely to be related to the outcome.

More than one third of the women dropped out of the trial:

- 17% dropped out after receiving only one injection of DMPA
- 13% dropped out after receiving only two injections of DMPA
- 7% dropped out after receiving three injections of DMPA

When the dropout rates are broken down by dose group, the rates were marginally higher in the 150 mg dose group.

Analytic Goal: Estimate dosage specific rates of amenorrhea that would have been observed in the absence of dropout and evaluate how sensitive inferences are to differing assumptions regarding dropout.

Notation

- N individuals observed at same set of occasions $\{t_1, t_2, ..., t_n\}$

- Let Y_{ij} denote the response for i^{th} individual at j^{th} occasion

- Y_i^c denotes the $n \times 1$ *complete* response vector, $Y_i^c = (Y_{i1}, ..., Y_{in})'$

- Let X_{ij} be a $p \times 1$ vector of covariates measured at t_j, $j = 1, \ldots, n$

- Let $X_i = (X_{i1}, ..., X_{in})'$ denote the matrix of covariates

- Primarily interested in making inferences about mean of $f(Y_i^c | X_i)$,
  e.g., $E(Y_i^c | X_i) = X_i\beta$ or $g[E(Y_i^c | X_i)] = X_i\beta$.

Dropout

- Each subject has a discrete event time D_i, denoting nonignorable dropout

- Let $D_i \in \{t_1, \ldots, t_n\}$ denote the last observed measurement occasion

- Dropout is "nonignorable" when D_i depends on unobserved Y_{ij}

- If $D_i \neq t_n$, ith subject is a "dropout"; otherwise, a "completer"

- Let $\phi_{ij} = \Pr(D_i = t_j)$

Observed Data

- Let Y_i denote the $n_i \times 1$ vector of the responses observed on the ith individual, i.e., the <u>observed</u> portion of Y_i^c

- *Observed* data for each subject consist of (Y_i, D_i, X_i)

- The covariates in X_i will generally include treatment or exposure group, in addition to time (t_j)

Models for Joint Distribution of (Y_i^c, D_i)

To correct for bias when dropout is nonignorable, joint models for the multivariate outcomes and dropout indicators have been proposed.

Little and Rubin (1987, 2002) and Little (1993; 1995) identified two broad classes of joint models:

1. Selection Models

2. Pattern Mixture Models

Selection Models

Joint distribution of Y_i^c and D_i is written as follows,

$$f(Y_i^c, D_i | X_i) = f_Y(Y_i^c | X_i) \, f_{D \cdot Y}(D_i | Y_i^c, X_i).$$

In longitudinal studies, primary focus is on inferences about $f_Y(Y_i^c | X_i)$.

$f_{D \cdot Y}(D_i | Y_i^c, X_i)$ plays the role of “nuisance parameters”, which can be ignored only if $f(D_i | Y_i^c, X_i)$ does not depend upon any missing Y_{ij}’s (or random effects).

Examples: Wu and Carroll (1988); Diggle and Kenward (1994); Molenberghs, Kenward and Lesaffre (1997); Ten Have *et al.* (1998, 2000).

Pattern Mixture Models

Joint distribution of Y_i^c and D_i is written as follows,

$$f(Y_i^c, D_i | X_i) = f_D(D_i | X_i) \, f_{Y \cdot D}(Y_i^c | D_i, X_i).$$

In longitudinal studies inferences about $f_{Y \cdot D}(Y_i^c | D_i, X_i)$ are <u>not</u> usually of main interest.

Rather, the primary interest is on inferences about $f_Y(Y_i^c | X_i)$, obtained by averaging over the distribution of D_i.

Examples: Wu and Bailey (1989); Follmann and Wu (1995); Little (1993, 1994); Hogan and Laird (1997).

Comment

Models for nonignorable dropout are fundamentally nonidentifiable.

Inference is possible only when unverifiable assumptions are made.

Inescapable fact that all methods for handling nonignorable dropout have to make some unverifiable assumptions.

In longitudinal studies, this problem is ameliorated somewhat by the fact that there is some information about the response before dropout.

However, recognizing that identification is driven by unverifiable assumptions, sensitivity analysis is warranted.

Selection versus Pattern Mixture Models

Selection Models:

- Target of inference: Model includes parameters of primary interest
- Easy to formulate hypotheses about dropout process
- Difficult to infer how assumptions on dropout process translate into assumptions about distribution of unobserved responses
- Difficult to determine model identifiability
- Computationally intractable

Pattern Mixture Models:

- Target of inference: Model excludes parameters of primary interest
- Make explicit assumptions about distribution of unobserved responses
- Implied dropout process is not immediately transparent
- Straightforward to determine model identifiability
- Computationally simple

Marginally-Specified Pattern Mixture Models

Recall: Basic idea underlying pattern mixture models,

$$f(Y_i^c, D_i|X_i) = f_D(D_i|X_i)\, f_{Y\cdot D}(Y_i^c|D_i, X_i),$$

is statification by different patterns of dropout.

Pattern mixture models for longitudinal data must incorporate dependence of Y_i^c on D_i as well as X_i.

That is, distribution of Y_i^c (given X_i) for those who dropout must be related to the distribution of Y_i^c for those who complete the study.

Example

Consider models for Y_{ij}, conditional on the time of dropout, that are of the following general form:

$$g\left[E\left(Y_{ij}|X_{ij},D_i\right)\right] = Z_{ij}'\beta^*$$

where g(·) is a known link function (e.g., log or logit), design vector Z_{ij} depends on dropout time, D_i and also incorporates the covariates X_{ij}.

Thus, conditional mean of Y_{ij} might depend on D_i and any other covariates (e.g., treatment or exposure group, time), and their interactions.

Note that the model for conditional mean of Y_{ij} will not be identified unless some (unverifiable) assumptions are made.

Recall: In a longitudinal study parameter of primary interest is <u>not</u> β^*.

Rather, the target of inference is the marginal expectation of the repeated outcomes,

$$E(Y_{ij}|X_{ij}) = \mu_{ij} = \sum_{l=1}^{n} \phi_{il}\, g^{-1}(Z'_{ij}\beta^*),$$

where Z_{ij} depends on the dropout patterns, and ϕ_{il} depends on X_i (or some subset of X_i).

Problem:

For non-linear link function, $g(\cdot)$, if

$$g\left[E(Y_{ij}|X_{ij}, D_i)\right] = Z'_{ij}\beta^*$$

then

$$g\left[E(Y_{ij}|X_{ij})\right] \neq X'_{ij}\beta$$

Illustration:

For example, if

$$\text{logit}\left[E\left(Y_{ij}|X_{ij},D_i\right)\right] = Z_{ij}'\beta^*$$

then

$$\begin{aligned}
\text{logit}\left[E\left(Y_{ij}|X_{ij}\right)\right] &= \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) \\
&= \log\left(\frac{\sum_{l=1}^{n}\phi_{il}\,\frac{\exp(Z_{ij}'\beta^*)}{1+\exp(Z_{ij}'\beta^*)}}{1-\sum_{l=1}^{n}\phi_{il}\,\frac{\exp(Z_{ij}'\beta^*)}{1+\exp(Z_{ij}'\beta^*)}}\right) \\
&\neq X_{ij}'\beta.
\end{aligned}$$

Marginally-Specified Pattern Mixture Models

To circumvent some of the problems with pattern mixture models, we propose marginally-specified models that involve three main components:

(i)	Marginal model for mean of Y_{ij}:	$E(Y_{ij}\|X_{ij})$
(ii)	Marginal model for dropout pattern, D_i:	$f_D(D_i\|X_i)$
(iii)	Conditional model for mean of Y_{ij} given D_i:	$E(Y_{ij}\|D_i, X_{ij})$

(i) Marginal model for mean of Y_{ij}:

$$g\left[E\left(Y_{ij}|X_{ij}\right)\right] = X_{ij}'\beta$$

(ii) Marginal model for D_i:

The multinomial probabilities for dropout, $\phi_i = (\phi_{i1}, \ldots, \phi_{in})'$, can simply be estimated as the sample proportion with each dropout time (stratified by exposure or treatment group and, perhaps, by other relevant covariates).

Alternatively, can consider parametric models for ϕ_i.

(iii) Conditional model for mean of Y_{ij} given D_i:

$$g\left[E(Y_{ij}|X_{ij},D_i)\right] = \Delta_{ij} + Z'_{ij}\beta^*$$

where Z_{ij} depends on D_i and also incorporates the covariates X_{ij}.

Note 1: Δ_{ij} is defined implicitly as a function of β, β^*, ϕ_i, since

$$E(Y_{ij}|X_{ij}) = \mu_{ij} = \sum_{l=1}^{n} \phi_{il}\, g^{-1}(\Delta_{ij} + Z'_{ij}\beta^*).$$

Note 2: (i), (ii), and (iii) specify a semi-parametric model.

Estimation of β

Once identifying constraints are adopted, β (and β^*) can be estimated via the solution to a set of GEEs:

$$\sum_{i=1}^{N} G_i' V_i^{-1} \left[Y_i - E(Y_i | X_i, D_i) \right] = 0,$$

where

$$G_i = \frac{\partial E(Y_i | X_i, D_i)}{\partial \theta}, \text{and } \theta = (\beta', \beta^{*\prime}),$$

and V_i is an appropriate weight matrix.

Note: Solution to GEE also requires solving for implicitly defined Δ_{ij}:

$$g^{-1}(X_{ij}'\beta) = E(Y_{ij} | X_{ij}) = \mu_{ij} = \sum_{l=1}^{n} \phi_{il}\, g^{-1}(\Delta_{ij} + Z_{ij}'\beta^*).$$

Concluding Remarks

Selection and pattern mixture models have their own distinct advantages and disadvantages.

Marginally-specified pattern mixture models capitalize on desirable features of each approach:

- marginally-specified pattern mixture models circumvent the obvious drawback of pattern mixture models
- by construction, regression parameters in marginally-specified pattern mixture models have “marginal” interpretations
- unlike selection models, identifiability restrictions are readily established
- estimation is relatively straightforward

The proposed model is semi-parametric.

The avoidance of full distributional assumptions can be advantageous:

- avoids having to make identifying restrictions on higher-order moments
- often no convenient specification of joint distribution when Y_{ij} are discrete

The general approach is closely related to "marginally-specified conditional models" developed for complete data (e.g., Fitzmaurice and Laird, 1993; Azzalini, 1994; Heagerty and Zeger, 2000).

Extensions to more general patterns of missing data are, in principle, straightforward.