Robust likelihood-based analysis of longitudinal data with missing values

Roderick Little and Guangyu Zhang
University of Michigan
Abstract

One approach to the analysis of longitudinal survey data with missing values is to multiply impute the missing values using draws from their predictive distribution given the observed data (Rubin 1987; Little and Rubin 2002, chapter 10). An advantage of the method is that predictors that are not relevant or not available for the final analysis model can be used in the imputation model. A robust model-based multiple imputation approach is proposed based on extensions of the work of Little and An (2004). Imputations are based on a model that models the relationship with the response propensity non-parametrically by a penalized spline, and models relationships with other variables parametrically. This approach has a form of double robustness that will be described, and simulation comparisons with other methods suggest that the method works well in a wide range of populations, with little loss of efficiency relative to parametric models when the latter are correct. Extensions to general patterns of missing data and the parameters other than unconditional means are outlined.

KEYWORDS: double robustness, incomplete data, penalized splines, regression imputation, weighting, longitudinal surveys

1. Preliminaries

Missing values arise in longitudinal studies for many reasons. A primary source is attrition, when subjects drop out prior to the end of the study. In longitudinal surveys, some individuals provide no information because of non-contact or refusal to respond (unit nonresponse). Other individuals are contacted and provide some information, but fail to answer some of the questions (item nonresponse). Often indices are constructed by summing values of particular items. For example, in economic studies, total net worth is a combination of values of individual assets or liabilities, some of which may be missing. If any of the items that form the index are missing, some procedure is needed to deal with the missing data.
The missing data pattern simply indicates which values in the data set are observed and which are missing. Specifically, let \( Y = (y_{ij}) \) denote an \((n \times p)\) rectangular dataset without missing values, with \( i \)th row \( y_i = (y_{i1}, \ldots, y_{ip}) \) where \( y_{ij} \) is the value of variable \( Y_j \) for subject \( i \).

In repeated measures problems variables measured at each time point form blocks of columns; for a single outcome \( Y_j \) denotes the value of a variable at the \( j \)th time point. With missing values, the pattern of missing data is defined by the missing-data indicator matrix \( M = (m_{ij}) \) with \( i \)th row \( m_i = (m_{i1}, \ldots, m_{ip}) \), such that \( m_{ij} = 1 \) if \( y_{ij} \) is missing and \( m_{ij} = 0 \) if \( y_{ij} \) is present. We assume throughout that \((y_{i}, m_i)\) are independent over \( i \).

The performance of alternative missing-data methods depends strongly on the missing-data mechanism, which concerns the reasons why values are missing, and in particular whether missingness depends on the values of variables in the data set. For example, subjects in a longitudinal intervention may more likely to drop out of a study because they feel the treatment was ineffective, which might be related to a poor value of an outcome measure. Rubin (1976) treated \( M \) as a random matrix, and characterized the missing-data mechanism by the conditional distribution of \( M \) given \( Y \), say \( f(M|Y, \phi) \), where \( \phi \) denotes unknown parameters. When missingness does not depend on the values of the data \( Y \), missing or observed, that is,

\[
\hat{f}(M|Y, \phi) = f(M|\phi) \text{ for all } Y, \phi,
\]

the data are called missing completely at random (MCAR). With the exception of planned missing-data designs, MCAR is a strong assumption, and missingness often does depend on recorded variables. Let \( Y_{\text{obs}} \) denote the observed values of \( Y \) and \( Y_{\text{mis}} \) the missing values. A less restrictive assumption is that missingness depends only on values \( Y_{\text{obs}} \) that are observed, and not on values \( Y_{\text{mis}} \) that are missing. That is,
The missing data mechanism is then called missing at random (MAR). Many methods for handling missing data assume the mechanism is MCAR or MAR, and yield biased estimates when the data are not MAR (NMAR).

2. Multiple Imputation for Repeated Measures Data

Two model-based approaches to repeated measures data with missing values are (a) to compute maximum likelihood (ML) or restricted ML (REML) estimates, based on the model, with standard errors based on the information matrix, the sandwich estimator, or the bootstrap (e.g. Little and Rubin 2002, chapter 6); or (b) to create a rectangular data matrix by multiply-imputing the missing values, analyzing the imputed data sets using a procedure for complete data, and then combining the estimates and standard errors using MI combining rules (Rubin 1987, 1996; Little and Rubin 2020, chapter 10). The ML approach is widely available in statistical software packages, for the case where the missing data mechanism is ignorable. Methods are available for normal models (SAS 1992; Pinheiro and Bates 2000) and for binary and ordinal repeated measures data ( ). Strengths of these methods are that they can handle a wide range of modeling situations, they are principled and have optimal asymptotic properties under the assumed model, and by constructing likelihoods based on the observed data, they avoid the need to impute the missing values. Weaknesses are that the inferences are asymptotic and potentially vulnerable to model misspecification. Also current widely-available software does not handle missing data in covariates or non-ignorable models (although the ML approach can be applied to handle such problems).

An alternative to these methods is to multiply-impute the missing values with draws from their predictive distribution under a model. MI allows the imputation uncertainty to be propagated in inferences by using MI combining rules. A principled version of this approach,
assuming MAR, is to impute using the posterior predictive distribution \( p(Y_{\text{mis}} \mid Y_{\text{obs}}) \) based on the model for \( Y \) and a prior distribution \( p(\theta) \) for the parameters \( \theta \) of this model. These imputations use the Bayesian paradigm to propagate uncertainty in the parameters. The imputations can be used for MI inferences, or can be viewed as a stepping stone to simulating directly the posterior distribution \( p(\theta \mid Y_{\text{obs}}) \) of the parameters, as in data augmentation (Tanner and Wong 1987; Schafer 1997; Little and Rubin 2002, chapter 10). When the imputations and complete-data inferences are based on the same model, say \( f(Y \mid \theta) \), and the prior distribution \( p(\theta) \) has support over the full parameter space of \( \theta \), this approach is asymptotically equivalent to the ML or REML approach mentioned above. This fact follows from the well-known large-sample equivalence of Bayes and ML, discussed for example in Little and Rubin (2002, chapter 6). The Bayes/MI approach can be a useful advance over ML for small samples; for example, in simple settings it recovers t corrections for normal models that are not a feature of the ML approach.

A useful feature of MI is that there is no requirement that the model for creating the imputations be the same as the analysis model applied to the filled-in data. This idea featured prominently in Rubin’s initial proposal of MI, which was in the context of imputation of public-use data sets where the imputer could not necessarily be expected to have the same analysis in mind as the analyst. A number of thorny theoretical issues arise when the imputation and analysis models differ and may not be mutually “congenial” (Fay 1996, Meng 1994, Rao 1996, Rubin 1987, 1996, Robins and Wang 2000), but the added flexibility opens up some attractive missing-data solutions for the practitioner. The resulting methods are not in all cases theoretically “pristine”, but they are likely to have excellent real-life properties. Here are some examples:

1. **Missing data in covariates.** As noted above, the ML approach as currently implemented in standard software does not allow for missing covariates. One possibility for dealing
with missing data in the covariates is to apply MI to impute them conditional on the value of $Y_{\text{obs}}$ and observed values of the covariates, using software such as IVEware (Raghunathan, Lepkowski, Van Hoewyk and Solenberger 2001). If there are not many missing covariate values, then a relatively simple imputation method that conditions on key predictors of the missing values should suffice. Then a repeated-measures model such as PROC MIXED can be applied to the imputed data sets, and results combined using MI combining rules. Multiple imputations of $Y_{\text{mis}}$ might be generated in the process of imputing the missing covariates, but these can be ignored for the PROC MIXED analysis, since it allows missing values in the $Y$'s.

2. **Auxiliary variables included in the imputation model.** Sometimes there are variables that are related to the missing variables and potentially related to the missing data mechanism that are not included in the final repeated-measures model. For example in a clinical trial, variables measuring side-effects of treatments may relate to drop-out and the outcome variables, but are not included in the final analysis model since they are post-treatment variables, and hence should not be included as covariates. Such variables might be included in the imputation model, but not in the final analysis model. Another example is the intention-to-treat analysis, where variables indicating which treatments are actually received can be included in the imputation model, even though they are not included in the primary analysis for comparing the treatments (Little and Yau 1996).

3. **Combining non-parametric complete-data analyses with parametric imputations.** Any imputations require assumptions about the predictive distribution of the missing values, and these can be explicitly formulated and assessed within a parametric modeling framework. On the other hand, some favor less parametric approaches to the final analysis, such as nonparametric tests based on ranks, and these are often specified as the
main analysis in a clinical trial. Methods of imputation for rank data are not highly
developed. A compromise is to base MI’s on a parametric model, and then allow the final
analysis of the multiply-imputed data to be nonparametric, with results from the tests
combined using MI combining rules. The advantage of this approach is that the
parametric assumptions are confined to the imputations, and violations of these
assumptions have much less impact than if the same model was applied for the analysis
of the filled-in data. This is particularly true if the fraction of missing information in the
missing values is small, since then MI model misspecification tends to have a minor
impact on the final analysis.

Model misspecification remains a concern for the imputation model, particularly when the
fraction of missing information is high and the sample size is large. We discuss here MI based on
a penalized spline of propensity prediction (PSPP), an approach that provides robustness to
misspecification of the mean structure in the model. We assume the missing data mechanism is
MAR, but make some comments about modifications of the approach for NMAR missing data in
the concluding section.

3. Robust MAR inference with a single missing outcome

We first consider the problem of imputing missing values of a single continuous outcome
variable \( Y_p \) given fully observed covariates \( X_1, \ldots, X_q \) and prior repeated measures \( Y_1, \ldots, Y_{p-1} \). MI
then requires a regression model for the conditional distribution of \( Y_p \) given \( X_1, \ldots, X_q \) and
\( Y_1, \ldots, Y_{p-1} \). A standard parametric approach is to assume a normal linear model, such as

\[
\begin{align*}
[Y_p \mid X_1, \ldots, X_q, Y_1, \ldots, Y_{p-1}, \beta, \sigma^2] &\sim N(\mu(X_1, \ldots, X_q, Y_1, \ldots, Y_{p-1}; \beta), \sigma^2) \\
\mu(X_1, \ldots, X_q, Y_1, \ldots, Y_{p-1}; \beta) &= \beta_0 + \sum_{j=1}^{q} \beta_j X_j + \sum_{k=1}^{p-1} \beta_{q+k} Y_k,
\end{align*}
\] (1)
where \( N(\mu, \sigma^2) \) denotes a normal distributions with mean \( \mu \) and variance \( \sigma^2 \). We consider an elaboration of this model that provides robustness against misspecification of the mean function \( \mu \). Other robust approaches, such as replacing the normal distribution by a longer-tailed distribution like the t (Lange, Little and Taylor 1989) can provide protection against outliers, but arguably misspecification of the mean structure is more important, at least in large samples. A standard accepted way of elaborating (1) is to add nonlinear terms and interactions to the model. With a single predictor, a more flexible approach is to model the relationship of that variable with \( Y_p \) via a spline function (e.g. Cheng 1994). This can be extended to handle a small number of covariates, but is subject to the “curse of dimensionality” when the number of covariates is large. The PSPP imputation model addresses this problem by confining the spline to a particularly important direction in covariate space, namely the direction defined by the propensity that \( Y_p \) is missing, and allowing a parsimonious form for other covariates.

Specifically, we define the logit of the propensity for \( Y_p \) to be observed, namely:

\[
Y_p^* = \logit \left( \Pr(M_p = 0 \mid X_1, \ldots, X_q, Y_1, \ldots, Y_{p-1}) \right),
\]

where \( M_p \) is the indicator for whether \( Y_p \) is missing. We can estimate \( Y_p^* \) by a logistic regression of \( M \) on \( (X_1, \ldots, X_q, Y_1, \ldots, Y_{p-1}) \), yielding an estimated propensity \( \hat{Y}_p^* \); this logistic regression can include nonlinear terms and interactions in \( (X_1, \ldots, X_q, Y_1, \ldots, Y_{p-1}) \), as necessary.

We then predict the missing values of \( Y_p \) using the following PSPP model:

\[
[Y_p | Y_p^*, Y_2, \ldots, Y_{p-1}, \beta, \phi, \sigma^2] \sim N \left( \mu(X_1, \ldots, X_q, Y_1, \ldots, Y_{p-1}), \sigma^2 \right),
\]

\[
\mu(X_1, \ldots, X_q, Y_1, \ldots, Y_{p-1}) = \hat{s}_p(Y_p^*; \phi) + g(X_1, \ldots, X_q, Y_1, Y_2, \ldots, Y_{p-2}, \beta)
\]

(3)
where \( s_p(\hat{Y}_p^*;\phi) \) is a penalized spline (discussed below), and \( g \) is a parametric function indexed by unknown fixed parameters \( \beta \), with the property that
\[
g(\hat{Y}_p^*,0,...,0,\beta) = 0 \quad \text{for all } \beta.
\] (4)

The variable \( Y_{p-1} \) is left out of the function \( g \) in (3) to avoid multicollinearity with \( \hat{Y}_p^* \). In general \( g \) should be chosen to include variables that are predictive of \( Y_p \) after conditioning on \( \hat{Y}_p^* \). A simple example of \( g \) that satisfies (4) is the linear additive model
\[
g(X_1,...,X_q,Y_1,Y_2,...,Y_{p-2},\beta) = \beta_0 + \sum_{j=1}^q \beta_j X_j + \sum_{j=q+1}^{p-2} \beta_{j+q} Y_j, \quad (5)
\]
but the model also allows interactions with \( Y_1^* \), for example:
\[
g(X_1,...,X_q,Y_1,Y_2,...,Y_{p-2},\beta) = \beta_0 + \sum_{j=1}^q \beta_j X_j + \sum_{j=q+1}^{p-2} \beta_{j+q} Y_j + \beta_{p+q-1} X_1 \hat{Y}_p^*.
\]

The spline \( s_p(\hat{Y}_p^*;\phi) \) could be based on any method that provides a flexible shape for the regression of \( Y_p \) and \( \hat{Y}_p^* \). We have used a penalized spline with a large number of fixed knots, based on the mixed model with a power-truncated spline basis (Eilers and Marx (1996); Ruppert and Carroll (2000)):
\[
s_p(\hat{Y}_p^*;\phi) = \phi_0 + \sum_{j=1}^q \phi_j \hat{Y}_p^{*j} + \sum_{k=1}^K \phi_{q+k} (\hat{Y}_p^* - \tau_k)_+^q, \quad (6)
\]
where \( q \) is the degree of polynomial, \((x)_+^q = x^q I(x \geq 0)\), \( \tau_1 < \cdots < \tau_K \) are selected fixed knots, and \( K \) is the total number of knots. The parameters \((\phi_0,...,\phi_q)^T\) in (6) are treated as a fixed (assigned as noninformative prior in Bayesian terms), and smoothing is achieved by treating \((\phi_{q+1},...,\phi_{q+K})^T\) as normal with mean zero and variance \( \lambda \sigma^2 \), where \( \lambda \) is a tuning parameter estimated from the data. Some aspects of this model are the following:
(1) The key idea of the model is that the relationship between $Y_p$ and $Y_p^*$ is modeled flexibly by a spline. It is important to get that relationship right, since misspecification of this aspect leads to bias. Conditional on $Y_p^*$, the distribution of the predictors is the same for respondents and nonrespondents, by the basic balancing property of the propensity score (Rosenbaum and Rubin 1983). This means that misspecification of the form of $g$ does not lead to bias for the mean of $Y_p$ given $Y_p^*$ (although it may lead to bias for other parameters). Thus for imputation, the form of $g$ is less important than the form of the regression on $Y_p^*$, although it is useful to include in $g$ predictors of $Y_p$ that improve the precision of the imputations. More generally, $g$ should include variables with relationships with $Y_p$ that are of interest in the final analysis of the imputed data.

(2) ML estimates $(\hat{\beta}, \hat{\phi}, \hat{\sigma}^2, \hat{\lambda})$ of the parameters $(\beta, \phi, \sigma^2, \lambda)$ can be obtained by ML or REML using widely available software such as PROC MIXED in SAS (SAS, 1992) and lme( ) in S-plus (Pinheiro and Bates 2000)). Missing values of $Y_p$ can then be drawn from their predictive distribution of $Y_p$ given $Y_{obs}$ and $(\hat{\beta}, \hat{\phi}, \hat{\sigma}^2, \hat{\lambda})$. This MI approach is improper (Rubin, 1987) in that it does not propagate uncertainty in estimating the parameters. A proper MI procedure is to compute ML estimates of each set of imputations from a bootstrapped sample of cases. Alternatively, a fully Bayesian approach is to draw values $(\hat{\beta}^{(d)}, \hat{\phi}^{(d)}, \hat{\sigma}^{2(d)}, \hat{\lambda}^{(d)})$ of the parameters from their posterior distribution using the Gibbs’ sampler, and then draw the missing values $Y_{mis}^{(d)}$ of $Y_p$ from their predictive distribution of $Y_p$ given $Y_{obs}$ and $(\hat{\beta}^{(d)}, \hat{\phi}^{(d)}, \hat{\sigma}^{2(d)}, \hat{\lambda}^{(d)})$. 

(3) One can think of $\hat{Y}^*_p$ as replacing one of the predictors in the model. Thus the function $g$ should usually exclude one of the predictors to avoid multicollinearity problems.

(4) Uncertainty in the logistic regression coefficients that determine $\hat{Y}^*_p$ can be ignored as a first approximation, or propagated by drawing them from their posterior distribution or computing them on a bootstrapped sample for each set of MI’s.

(5) The “double robustness” property is that the estimated mean of $Y_p$ for this model is consistent if either (a) the function $\mu$ is correctly specified, or (b) the response propensity is correctly specified and the spline correctly models the relationship between $Y_p$ and the response propensity. In the original formulation of the model (Little and An 2004), this “double robustness” property was shown to hold when the variables in the parametric function $g$ were centered by replacing them by residuals from regressions on $\hat{Y}^*_p$. Subsequently, Zhang and Little (2005) showed that the property still holds for the simpler version (3) of the model that omits this centering.

(6) Another way to achieve robustness for the mean of $Y_p$ is to model the relationship between $Y_p$ and the predictors parametrically, and calibrate the predictions from the model by adding a weighted mean of the residuals, with weights that are the inverse of the response propensity. This leads to a “calibration” estimator of the form

$$\hat{\mu}_p = n^{-1}\left(\sum_{i=1}^n w_i (y_{ip} - \hat{y}_{ip})\right) + n^{-1}\left(\sum_{i=1}^n \hat{y}_{ip}\right).$$

(7) where $\hat{y}_{ip}$ is the prediction for case $i$. The estimator (7) has properties of semi-parametric efficiency and “double-robustness” (Robins, Rotnitsky and Zhao (1994); Robins and Rotnitsky (2001), Scharfstein and Irizarry 2003), in the sense that the estimate is consistent if just one of
the models for prediction and weighting are correctly specified. However, since the calibration of the predictions is to correct effects of model misspecification, we believe that the calibration of the predictions (7) is unnecessary if the prediction model does not make strong parametric assumptions, as in (4). This conjecture is supported by the simulation studies in Little and An (2004), where the PSPP method has similar performance to the calibration approach and dominates it for some choices of simulation parameters.

4. Extensions of PPSP to Monotone and General Patterns

For longitudinal data subject to attrition, the resulting data have a monotone missing data pattern for the repeated measures. The PPSP approach of the previous section can be applied sequentially to create MI data sets for a monotone pattern by fitting a model of form (3) to the distribution of $Y_j$ given $(X_1,\ldots,X_q,Y_1,\ldots,Y_{j-1})$, with the response propensity for $Y_j$ based on a logistic regression of its missing data indicator $M_j$ on $(X_1,\ldots,X_q,Y_1,\ldots,Y_{j-1})$ ($j = 1,\ldots,p$).

Missing values of $Y_j$ are replaced by draws from their predictive distribution and used to predict missing values of later values of $Y$. Missing values of covariates are replaced by their predictions in this sequence of regressions. For a general pattern of missing data, the sequential imputation methods of Raghunathan et al. (2001) can be extended to provide PPSP imputations that condition on the spline of the propensity that each variable is missing. An (2005) discusses these extensions in detail.

5. Extensions to inferences other than means

The PSPP method provides robustness for estimating the marginal mean of $Y$, but this robustness property does not necessarily apply for other parameters, such as subgroup means. In particular in longitudinal clinical trials, we are often primarily concerned with comparisons of means between treatment groups, which are contrasts of means in subgroups defined by the
treatment variable. Zhang and Little (2005) show in simulations that the PSPP method does not provide estimates of treatment contrasts that are protected against model misspecification, even if the treatment variable is included as a covariate in the $g$ function. Zhang and Little (2005) propose extensions of the PSPP, stratified PSPP and bivariate PSPP, to address this issue.

Specifically, suppose that missing values are confined to $Y_p$, and interest is in the mean of $Y_p$ given $X_1$, where $X_1$ is a categorical variable with $C$ categories. The stratified PSPP method computes a separate spline on the propensity to respond in each of the categories of $X_1$, as follows:

(a) estimate the response propensity

$$Y_p^* = \logit \left( \Pr(M_p = 0 \mid X_1, X_2, \ldots, X_q, Y_1, \ldots, Y_{p-1}) \right)$$

(b) Let $I_c = 1$ if $X_1 = c$; $I_c = 0$ if $X_1 \neq c$, $c = 1, \ldots, C$, and form the propensity in each category of $X_1$:

$$Y_p^* = Y^* \times I_c, \quad c = 1, \ldots, C.$$ Then model the conditional distribution of $Y_p$ as in (3) with mean in category $X_1 = c$

$$\mu(X_1 = c, X_2, \ldots, X_q, Y_1, \ldots, Y_{p-1}) = s_{pc}(\hat{Y}_p^*; \phi) + g(X_1, \ldots, X_q, Y_1, Y_2, \ldots, Y_{p-1}, \beta)$$

where $s_{pc}(Y^*)$, $c = 1, \ldots, C$, is a spline for the regression of $Y$ on $Y^*$ and $g$ is as in (3).

This method yields a doubly robust property for the conditional mean of $Y$ given $X_1$ (Zhang and Little 2005). The marginal mean of $Y$ is a weighted average of conditional means, which again has the consistency property.

Zhang and Little (2005) also consider estimating the conditional mean of $Y_p$ given a continuous variable $X_1$, assuming a linear regression function

$$E(Y_p \mid X_1) = \gamma_0 + \gamma_1 X_1.$$ To achieve
the double robustness property for estimates of $\gamma_0, \gamma_1$, we need to replace the univariate spline on the propensity by a bivariate spline on the propensity and $X_1$.

6. Example

To illustrate these methods we use data from an online weight loss study conducted by Kaiser Permanente, the analysis of which is discussed in Couper, Peytchev, Little, Strecher and Rothert (2005). Approximately 4,000 subjects were randomly assigned to the treatment or the control group. For the treatment group, the weight loss information provided online was tailored to the subjects based on their answers to an initial survey, which contained baseline measurements such as weight, motivation to weight loss, etc; for the control group, information provided online was the same for all the subjects. At 3 months, a second survey was sent to all of the participants. Our goal is to compare the short-term treatment effects; in particular, we compare the reduction of the body mass index (BMI).

There were 2059 subjects in the treatment group and 1956 subjects in the control group at the baseline. At 3 months, 623 subjects in the treatment group and 611 subjects in the control group responded to the second survey. We assume the data are missing at random. Comparisons of the baseline measurements between subjects remained in the study and those dropped out indicated that those who remained in the study were older than those who dropped out of the study ($P=0.01$ for the treatment group and $P<0.01$ for the control group); for the treatment group, subjects who remained in the study have much lower baseline BMI than those who dropped out of the study ($P<0.001$), but this difference is not seen in the control group ($P=0.47$), suggesting an interaction between treatment and baseline BMI that is included when estimating the propensity scores.

We applied the PSPP method and the stratified PSPP method to the data and Table 1 shows the results. The treatment group has a larger reduction of BMI after 3 month compared to
the control group (P<0.01) based on the complete case analysis. The stratified PSPP method and the PSPP method with the treatment as a covariate yield similar results. On the other hand, the PSPP method without treatment as a covariate does not show the treatment effect (P=0.21). We fit an ANCOVA model with all the covariates included in the stratified PSPP method. It yields similar results (P<0.01).

In this example, we find the reduction of BMI is linear with respect to the propensity score. As a result, the stratified PSPP method and the PSPP method with the treatment as a covariate yield similar results. And the PSPP methods show no difference from the parametric method.

Table 1 BMI reduction within groups

<table>
<thead>
<tr>
<th>Method</th>
<th>Treatment</th>
<th>Control</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
</tr>
<tr>
<td>CC</td>
<td>-0.92</td>
<td>0.09</td>
<td>-0.44</td>
</tr>
<tr>
<td>Stratified PSPP</td>
<td>-0.99</td>
<td>0.11</td>
<td>-0.42</td>
</tr>
<tr>
<td>PSPP with treatment as a covariate</td>
<td>-0.96</td>
<td>0.11</td>
<td>-0.45</td>
</tr>
<tr>
<td>PSPP without treatment as a covariate</td>
<td>-0.79</td>
<td>0.09</td>
<td>-0.63</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>-0.98</td>
<td>0.10</td>
<td>-0.44</td>
</tr>
</tbody>
</table>

*SE is based on 200 bootstrap samples.

7. Discussion

We have described an MI approach to the analysis of longitudinal data with missing values, which provides flexibility by allowing the analysis model and the imputation model to differ. A good MI model should make relatively weak assumptions about the predictive distribution of the missing values, particularly when samples are large and missing data are
extensive. In this vein, we have discussed MI models that include a spline on the propensity to be missing, which provide a form of double robustness and hence some protection against misspecification of the mean structure. We have considered normal models here, but for non-normal outcomes the same approach to the mean structure can be applied to generalized linear models, with the systematic component of the link function replacing the mean.

Throughout the chapter we have assumed that the missing data are MAR, so the question remains how to handle situations where the mechanism is thought to be NMAR. The model-based MI approach to missing data can be applied to NMAR models by formulating a model for the joint distribution of the data and the missing data mechanism (e.g. Diggle and Kenward 1994; Little 1995; Little and Rubin 2002, chapter 15). However, lack of identifiability of the parameters is a serious issue in fitting these models, as deviations from MAR cannot be detected from the data without making strong structural or distributional assumptions. The impact of NMAR nonresponse can be reduced by the following strategies:

(a) At the design stage, record values of covariates that are predictive of nonresponse, and condition on these in imputing the missing values.

(b) Consider following up a subsample of nonrespondents to recover at least the key missing information on these cases. These data can then be used to multiply-impute the information on incomplete cases that are not followed up.

(c) When there are various mechanisms of missingness, attempt to determine which of the missing values of a variable are likely MAR and which are likely not MAR, and then use MAR methods to multiply impute the former. This reduces the scope of the NMAR problem for this variable, as compared with an analysis that fits an NMAR model to all the missing values.

For the missing values that are thought to be NMAR, we recommend a sensitivity analysis under plausible alternative models for nonignorable nonresponse (Rubin (1977); Little
and Wang (1996); Scharfstein, Rotnitsky and Robins (1999)). As exemplified in Rubin (1977), the Bayesian framework is convenient for this modeling, since it explicitly recognizes the need for prior information. A simple approach, which may be as effective as more complex alternatives, is to introduce offsets into the means of the predictive distributions used to multiply-impute the missing values that are thought to be NMAR. For example, in the sequential IVEware approach of Raghunathan et al. (2001), imputations for a missing variable $Y_j$ at a particular iteration $t$ are based on the predictive distribution of $Y_j$ given the other variables, with missing values of the other variables imputed with draws from previous steps of the sequence. Suppose some set $\{y_{\vec{i}}, i = 1, \ldots, r\}$ of the missing values of $Y_j$ are thought to be NMAR, and let

\[
\{\mu_j^{(t)} : i = 1, \ldots, r\} \text{ and } \{\sigma_j^{(t)} : i = 1, \ldots, r\}
\]

be the means and standard deviations of the predictive distributions of these values at iteration $t$ under an MAR model. One possible sensitivity analysis consists of replacing these predictive means by $\{\mu_j^{(t)} + \delta_j \sigma_j^{(t)} : i = 1, \ldots, r\}$ for one or more values of $\delta_j$ that are thought to represent plausible deviations from MAR, for example $\delta_j = 0.5$ if a positive deviation of one half a standard deviation is reasonable, $\delta_j = -0.5$ if a negative deviation of that size is thought reasonable, or both $\delta_j = 0.5$ and $\delta_j = -0.5$ if the sign of the deviation is unknown. This change is easily accomplished by replacing the MAR imputations $\{\hat{y}_j^{(t)} : i = 1, \ldots, r\}$ by NMAR imputations $\{\hat{y}_j^{(t)} + \delta_j \sigma_j^{(t)} : i = 1, \ldots, r\}$. What this method lacks in sophistication may be compensated by the transparency of the underlying assumptions. A Bayesian variant is to assign $\delta_j$ a prior distribution and draw values of that parameter at each iteration.
Acknowledgments

This research was supported by
References


